Winter 12-22-2009

Evolution of Genome-wide Gene Regulation in the Budding Yeast Cell-Division Cycle

Daniel F. Simola
University of Pennsylvania, simola@mail.med.upenn.edu

Evolution of Genome-wide Gene Regulation in the Budding Yeast Cell-Division Cycle

Abstract
Genome-wide regulation of gene expression involves a dynamic epigenetic structure which generates an organism's life-cycle. Although changes in gene expression during development have broad effects on many basic phenomena including cell growth, differentiation, morphogenesis, and disease progression, the evolutionary forces influencing gene expression dynamics and gene regulation remain largely unknown, due to the nature of gene expression as a polygenic, quantitative trait. Moreover, gene expression is regulated differentially over time, so evolutionary forces may be influenced by developmental context. To advance the understanding of evolution in the context of the life-cycle, the architecture of gene expression timing control and its influence on expression dynamics must be revealed. This dissertation presents two experimental investigations of the evolution of genes and related structural regions and time-dependent gene expression, using the budding yeasts Saccharomyces cerevisiae and Saccharomyces paradoxus and their mitotic cell-division cycle as model organism and life-cycle. Comparative methodologies were employed to analyze genome-wide patterns of genetic and phenotypic diversity within and between species. Analysis of several dozen yeast genomes reveals a dominant evolutionary mode of purifying selection. Despite limited genetic variability, differences in transcriptional regulation appear to contribute predominantly to interspecies divergence, and altered post-transcriptional regulation of ribosomal genes may have altered the timing of each species' transition from vegetative growth to reproduction, a classic life-history trait. In addition, natural variation in genome-wide gene expression was measured as a time-series through the mitotic cell-division cycle of 10 yeast lines, including one outgroup species. Despite levels of variation consistent with strong stabilizing selection, transcriptome coexpression dynamics have diverged significantly within and between species. A model involving timing pattern changes explains 61% of the between-genome variation in expression dynamics, suggesting that the major mode of transcriptome evolution involves changes in timing (heterochrony) rather than changes in levels (heterometry) of expression. Analysis of heterochrony patterns suggests that timing control is organized into distinct and dynamically-autonomous modules. Divergence in expression dynamics may be explained by pleiotropic changes in modular timing control. Genome-wide gene regulation may utilize a general architecture comprised of multiple discrete event timelines, whose superposition could produce combinatorial complexity in timing patterns.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Junhyong Kim

Keywords
yeast, genomics, evolution, dynamics, heterochrony, modularity

Subject Categories
Computational Biology | Evolution | Genomics | Systems Biology

EVOLUTION OF GENOME-WIDE GENE REGULATION IN THE BUDDING YEAST CELL-DIVISION CYCLE

Daniel F. Simola

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2009

Junhyong Kim, Dissertation Supervisor

Maja Bucan, Graduate Group Chair

Dissertation Committee

Mark Goulian, Ph.D. (Committee Chair)

Wei Guo, Ph.D.

Vijay Kumar, Ph.D.

Paul D. Sniegowski, Ph.D.

Evolution of Genome-wide Gene Regulation

in the Budding Yeast Cell-Division Cycle

Daniel Francis Simola

To Melissa, whose loving patience sustained me in this journey

Acknowledgements

As someone whose habit it is to think in a breadth-first rather than depth-first manner, perhaps the most difficult part of graduate school for me was learning how to translate my thoughts, skills, and interests into specific aims. During the past six years many people have assisted me in this regard by offering their encouragement, advice, and support. Foremost among these are the members of my thesis committee—Mark Goulian, Wei Guo, Vijay Kumar, Paul Sniegowski, and Junhyong Kim—who oversaw my progress throughout graduate school, from initially formulating a coherent research project to eventually graduating. Through their varying research interests, they have offered constructive criticism and helpful advice. Moreover, their enthusiasm for science provided a refreshing influence on my work. Mark, thank you for being a willing and able committee chair. Paul, thank you for your discussions about yeast evolution when I first began research. Wei and Vijay, thank you for providing your encouragement, attention, and time from your busy schedules. Junhyong, you have delivered nothing short of an exceptional mentoring experience. More than anything I thank you for your patience and for providing ample room for intellectual growth, the chance to explore new concepts and various research interests, and a flexible work environment. You taught me how to focus on the details, see the big picture, and then communicate these insights to others. I have never worked with anyone so able to balance responsibilities, and I will be able to point to your guidance and example in any professional success I might have in the future.

I would also like to thank the faculty who taught me in various regards. Maja Bucan and Jim Eberwine provided great mentoring during laboratory rotations; Maja jump-started my experience working with gene expression data, and Jim provided cutting-edge opportunities to develop my wet lab skills. They both offered helpful advice during prelims. Warren Ewens, Sridhar Hannenhalli, and Sampath Kannan offered their time and talents guiding me through independent studies.

Chantal, Dexter, Fan, Hoa, Li-San, Miler, Sheng, Shreedhar, and Stephen have all contributed to making our lab a comfortable second home. Chantal's bright smile boosts the afternoon mood, and her company and skills in the wet lab made working more enjoyable. Thank you especially for your initial efforts working with our yeast strains; without them this dissertation would not have been possible. Dexter, Fan, and Sheng led the way towards graduation and showed me how to remain persistent to the end. Li-San brought great energy and enthusiasm to discussions and presentations in the lab. Hoa and Shreedhar have provided useful insights, advice, and help in the lab. I would have run into many hurdles if it hadn't been for Stephen, who has adeptly managed all of our computer and electronics equipment and any other lab necessities, while tackling his own research projects and another job. He has been incredibly responsive to my questions and needs and is a remarkable asset to the lab. Miler and I have been like "GCB twins" throughout graduate school; we started the program together, joined Junhyong's lab at the same time, got stuck sharing a windowless concrete closest in Goddard, collaborated on research, and are finally graduating together. Miler, thank you for your companionship, not to mention your many edible contributions to the lab.

Miler, Praveen, and Tom have been great friends throughout. They helped take the stress out of graduate school, helped me pass through various hurdles, and have provided great discussions, scientific and otherwise. I have no doubt you will all achieve great things.

Helen taught me proper PCR technique and helped generate gene expression data for the mutation accumulation lines. Cara helped me develop protocols for processing microarrays during her rotation in the lab. Overall, it has been a pleasure working with everyone in the GCB program. Your dedication and drive are what makes this program great.

One of the first things I realized after beginning research was how long it might take to get to this point. Junhyong has commented that graduate school is essentially an endurance test to determine, more than anything else, whether you can follow a question to its conclusion, while constantly faced with the slow progression of research and the fast pace of everyday life. Although this challenge obviously applies to students, it inevitably confronts family as well. I am very grateful to my long-time friend and wife who has been incredibly receptive of this challenge. Despite my false claims of "finishing soon", Melissa has wholeheartedly supported me throughout graduate school with an impressive degree of patience. I'm not sure how it might have turned out otherwise, but I cannot begin to thank her enough for how it did. In addition, my parents, Frank and Veronica, have served as a constant source of support throughout my life; they taught me how to pursue my interests with hard work.

Thank you, all.

ABSTRACT

EVOLUTION OF GENOME-WIDE GENE REGULATION
IN THE BUDDING YEAST CELL-DIVISION CYCLE

Daniel F. Simola

Junhyong Kim

Genome-wide regulation of gene expression involves a dynamic epigenetic structure which generates an organism's life-cycle. Although changes in gene expression during development have broad effects on many basic phenomena including cell growth, differentiation, morphogenesis, and disease progression, the evolutionary forces influencing gene expression dynamics and gene regulation remain largely unknown, due to the nature of gene expression as a polygenic, quantitative trait. Moreover, gene expression is regulated differentially over time, so evolutionary forces may be influenced by developmental context. To advance the understanding of evolution in the context of the life-cycle, the architecture of gene expression timing control and its influence on expression dynamics must be revealed. This dissertation presents two experimental investigations of the evolution of genes and related structural regions and time-dependent gene expression, using the budding yeasts *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* and their mitotic cell-division

cycle as model organism and life-cycle. Comparative methodologies were employed to analyze genome-wide patterns of genetic and phenotypic diversity within and between species. Analysis of several dozen yeast genomes reveals a dominant evolutionary mode of purifying selection. Despite limited genetic variability, differences in transcriptional regulation appear to contribute predominantly to interspecies divergence, and altered post-transcriptional regulation of ribosomal genes may have altered the timing of each species' transition from vegetative growth to reproduction, a classic life-history trait. In addition, natural variation in genome-wide gene expression was measured as a time-series through the mitotic cell-division cycle of 10 yeast lines, including one outgroup species. Despite levels of variation consistent with strong stabilizing selection, transcriptome coexpression dynamics have diverged significantly within and between species. A model involving timing pattern changes explains 61% of the between-genome variation in expression dynamics, suggesting that the major mode of transcriptome evolution involves changes in timing (heterochrony) rather than changes in levels (heterometry) of expression. Analysis of heterochrony patterns suggests that timing control is organized into distinct and dynamically-autonomous modules. Divergence in expression dynamics may be explained by pleiotropic changes in modular timing control. Genome-wide gene regulation may utilize a general architecture comprised of multiple discrete event timelines, whose superposition could produce combinatorial complexity in timing patterns.

Table of Contents

1 Introduction **1**

1.1 Background **3**
1.1.1 Yeast and its cell-division cycle as model organism and life-cycle . 3
1.1.2 Population structure and evolutionary history of *Saccharomyces* . . 5
1.1.3 Genome-scale molecular evolution in yeast 6
1.1.4 Gene expression 8
1.1.5 Evolution of genome-wide gene expression 9
1.1.6 Life-cycle evolution and heterochrony 12

1.2 Overview of Dissertation **13**

2 Evolutionary Dynamics of the *Saccharomyces* Genome and a Resource for Population Genomics **27**

2.1 Abstract **27**

2.2 Introduction **28**

2.3 Results **30**
2.3.1 Identification and partitioning of protein-coding loci 30
2.3.2 Heterozygosity within GSR categories 32
2.3.3 Evolution of GSR categories within species 33
2.3.4 Polymorphism and divergence of GSR categories 36

2.4 Discussion **43**
2.4.1 Confounding selection with demographics 45
2.4.2 Recombination and linkage disequilibrium 46
2.4.3 Population parameter estimates using neutral loci 47

2.5 Materials and Methods **49**

3 Heterochronic Evolution Reveals Modular Timing Changes in Budding Yeast Transcriptomes **74**

3.1 Abstract **74**

3.2 Introduction **75**

3.3 Results **76**

3.3.1 Genome-wide expression levels show much less variability than expected, but CDC-temporal expression patterns display broad divergence 76

3.3.2 CDC regulatory architecture exhibits time-dependent changes in multi-dimensional complexity 79

3.3.3 Divergence in coregulatory patterns is found at all scales of transcriptome organization 84

3.3.4 A hypothesis for modular timing control 87

3.3.5 Heterochrony explains evolution of expression dynamics 88

3.3.6 Shared patterns of heterochrony reveal modular timing changes . . 89

3.3.7 A modular, pleiotropic regulatory architecture explains CDC transcriptome divergence 100

3.4 Discussion **104**

3.5 Materials and Methods **106**

4 Conclusions and Future Directions **211**

4.1 Future Directions **215**

4.2 Conclusion **218**

Works Cited **241**

List of Tables

1.1 Genetic variability in proximal promoters of transcription factor loci among woodland *S. cerevisiae* isolates 17

2.1 Differences in estimates of site heterozygosity and Tajima's D between species 56
2.2 Genes with multiple GSRs under balancing selection 57
2.3 HKA statistics for each GSR category 58
2.4 Number of GO terms rejected by the HKA test within each GSR category . 59
2.5 Number of GSR categories rejected by the HKA test, grouped by GO term . 60
2.6 Polymorphism statistics for GSR categories 61
2.7 Divergence statistics for GSR categories 62
2.8 Controls for the genome identification algorithm 63
2.9 *S. cerevisiae* polymorphism statistics using genes common between species 65
2.10 Polymorphism statistics for transposable elements in *S. cerevisiae* 67

3.1 Natural and laboratory budding yeast isolates used in this study 185
3.2 Ranking of GO terms by proportion of associated genes evolving under stabilizing selection 186
3.3 Genes with neutral or partly neutral expression trajectories for each life-cycle related term 188
3.4 Estimates time-dependent transcriptome coexpression structure 190
3.5 Comparison of evolutionary covariance matrices across sequential CDC timepoints using the common principle components test 192
3.6 Gene enrichments and top 5 genes correlating with major and rank-2 eigengenes throughout the CDC 193
3.7 Gene enrichment of 88 GO-Slim terms for the top 5% of genes correlating with major eigengenes throughout the CDC 195
3.8 Ranking of gene groups by total number of genes across the CDC associated with a particular group 197
3.9 Statistics describing evolutionary divergence of the modular yeast coexpression structure 199
3.10 Top 50 heterochronic genes, ranked by timing pattern distortion 200

3.11 Enrichment of functional ontology terms in timing modules, using the set of 1828 within-module genes 202
3.12 Modular expression timeline variability for 7 timing modules 204
3.13 Heterochronic evolution in module-specific transcription factors 205

List of Figures

1.1 Variability in CDC length and cell size in late G_1-phase for 10 *Saccharomyces* isolates 16

2.1 Genome-wide distribution of $\hat{\pi}$, $\hat{\theta}$, and Tajima's D across GSR categories . 54
2.2 Scatterplot of number of ORFs recovered by the genome identification algorithm versus the whole-genome shotgun coverage of each genome 55

3.1 Overview of experimental design, analysis, results, and conclusion. 135
3.2 Genome-wide evolutionary gene expression variability among *S. cerevisiae* strains through the CDC 136
3.3 Relationship of two components of gene expression variation, time (temporal variation) and strain (strain divergence) 138
3.4 Hierarchical clustering of the entire CDC data set 139
3.5 Profiles for the 40 GO-Slim terms which exhibit the lowest average proportion of genes under stabilizing selection 140
3.6 Visualization of CDC-transcriptomes using 2-dimensional SVD projections 141
3.7 Comparison of cumulative eigenvalue distributions for MA line data (23 samples) and for *S. cerevisiae* CDC data at each timepoint (9 samples each) 143
3.8 Cumulative eigenvalue distributions for natural CDC-transcriptomes and MA lines 144
3.9 Comparison of singular value distributions for the *S. cerevisiae* CDC data (162 samples) and MA line data (23 samples) 146
3.10 Comparison of transcriptome cumulative eigenvalue distributions 147
3.11 Multivariate evolutionary transcriptome variability across the CDC for each of the top 8 CDC-directions 148
3.12 Temporal variability in CDC-transcriptome evolutionary covariance structure. 150
3.13 Angular distance matrices relating pairs of CDC-directions of the same rank 151
3.14 Example projection of expression data onto the top three global eigenvectors of *S. cerevisiae* 153
3.15 Heat maps illustrating Mantel matrix correlations between CDC-expression correlation matrices of natural strains 154
3.16 Evolutionary divergence of CDC-transcriptome coexpression structure within and between CDC-phase groups 155

3.17 Evolutionary divergence of major directions of covariation within CDC-phase groups . . . 156
3.18 Expected and observed proportions of k-module overlap for increasing numbers of genes k . . . 158
3.19 Heat maps and hierarchical clusterings indicating proportions of overlap between gene modules across strains . . . 159
3.20 Results of the heterochrony regression model explaining time-dependent changes in gene expression trajectories between strains . . . 161
3.21 Frequency distributions of optimal time-domain parameters . . . 163
3.22 Distributions of 95%-equivalent timing curve ensemble sizes and distances between ensembles . . . 165
3.23 Heat maps and timepoint plots illustrating distances between pairs of timing curves . . . 167
3.24 Visualization of timing curve distances by MDS . . . 169
3.25 Linear discriminant distance matrix visualization and timing curves for each cluster. . . . 170
3.26 Gap statistic and bootstrap analyses of timing pattern clusters . . . 171
3.27 2-dimensional linear discriminant plots of the distance relationships between timing curves for 4998 heterochronic genes for each of the 45 strain comparisons . . . 173
3.28 Distributions of co-cluster similarity between pairs of genes . . . 174
3.29 Modularity in the heterochronic gene interaction graph . . . 175
3.30 The modular architecture of genome-wide timing control . . . 177
3.31 Linear discriminant visualization of 7 timing pattern clusters . . . 179
3.32 Heat map of modular expression timeline variance . . . 180
3.33 Gene expression timeline variability across the CDC for 1828 timing module genes . . . 181
3.34 CDC-expression trajectories of genes involved in the timing of the G_2/M-phase transition . . . 182
3.35 Comparison of sequence divergence and expression divergence between *S. cerevisiae* and *S. paradoxus* . . . 183
3.36 Budding index profiles for natural *S. cerevisiae* strains. . . . 184

Chapter 1

Introduction

"Only a thorough-going study of variation will lighten our darkness."

J.B.S. Haldane, 1932

During the last few centuries biology has matured from a science defined primarily by characterizations of morphological (*1, 2*), embryological (*3, 4*), molecular (*5*), and genetic (*6*) patterns of variation to one at the crux of a complete evolutionary theory of bio-generative processes (*7, 8, 9, 10*). Simply characterizing and comparing the many patterns of variation in a cell seems a sufficiently great task, one that Robert Hooke began 350 years ago with his *Micrographia* (*11*), and one that remains far from complete. But to understand the true diversity of life we must not only systematically consider and compare these patterns, but use them to discern the underlying evolutionary and developmental processes responsible for their generation. Understanding this problem of variation must begin by uncovering the complexity inherent in life's most fundamental unit, the cell, together with its generative process, the life-cycle.

The life-cycle is a dynamic process involving growth, reproduction, aging, and death (*12, 13*) and relates a cell's development to its genome and evolutionary history (*14*). While genomes provide the encoding for life-cycles, expression of this genetic information into phenotype during development is necessary for evolution to occur (*15, 16*). Phenotype, the

collection of expressed traits resulting from a genome's time-dependent interactions with its environment, acts as an integrative filter through which heritable differences pass before being propagated or extinguished during evolution (*17*). Changes in the processes that generate phenotype can not only affect the phenotype itself, but also influence the evolution of the genotype, such that developmental dynamics give rise to evolutionary dynamics in the timescale where changes in the life-cycle become encoded into an organism's genome. In this way a population's life-cycle evolves, propagating not only genotype but an entire system of development (*18, 19*). Consequently, progress towards a "diachronic" evolutionary theory (*20*) requires investigations into the architecture of phenotypic expression in the context of the life-cycle.

This integrated approach was first advocated in the early twentieth century by Richard Goldschmidt, who combined both experimental genetics and comparative embryology to investigate the role of hereditary factors and their evolution on the expression of cellular phenotype (*21*). Despite ignorance of the molecular details of development, Goldschmidt emphasized the quantitative and dynamical nature of genes and argued that differences in the rates of gene expression, rather than qualitative differences in the genes themselves, contributed substantially to population divergence and speciation (*22*). Among Goldschmidt's contemporaries, Conrad Waddington helped to formalize this notion into a theory of evolution of developmental systems founded on genes (*23*), arguing that the expression and interaction of genes forms a complex, dynamic, and epigenetic process which instantiates life-cycles from genomes. Since time-dependent gene expression contributes to life-cycle progression, changes in the expression of and interactions among genes must contribute to life-cycle variation. Characterizations of the evolution of genes, their time-dependent expression, and their interactions are thus needed to understand the cellular basis for life-cycle generation and diversity.

1.1 Background

1.1.1 Yeast and its cell-division cycle as model organism and life-cycle

The term yeast refers to any unicellular eukaryotic fungi, however in this dissertation yeast refers specifically to organisms of the genus *Saccharomyces*, either *Saccharomyces cerevisiae* (commonly called budding or baker's yeast) or *Saccharomyces paradoxus* (a natural yeast species). The life history of these yeasts has provided characteristics that greatly facilitate biological research. Yeast has cells that are small (2–6 μM) but easily visible under a microscope, a short ultradian reproductive cycle, haploid and diploid life-cycle phases (haplodiplontic), sexual and asexual reproductive modes with two mating types, simple gene structure, and a compact genome (*24, 25*). They are easily cultured in a laboratory, can be frozen and revived for long-term storage, and can metabolize energy using either respiration or fermentation. This suite of characteristics has led to the adoption of yeast as a key model organism for studying genetics (*26*), the cell-division cycle (*27*), functional genomics (*28*), genome evolution (*29*), and most recently, ecology and evolution (reviewed in (*30, 31*)).

The yeast cell-division cycle (CDC) is typical of eukaryotes, comprising four stages or phases of progression: growth (G_1), DNA replication (S), second growth (G_2), and mitosis (M). These phases are separated by molecular checkpoints that monitor successful completion of each phase (*32*). In this way, the yeast CDC essentially consists of one long stage of growth involving duplication of all molecular components, followed by reproduction by partitioning these components into two cells. This is the minimum required for a life-cycle (*33*). Moreover, a cell can sense and coordinate various cellular and environmental signals to adjust its growth rate, progression through the CDC, and the method and timing of reproduction (*34*). This coordination is achieved not only through metabolic kinetics, but

through changes in the regulation of gene expression (*35, 36*), a necessary component of developmental progression (*37*), adaptation (*38*), and life history evolution (*39*). For these reasons I consider the yeast cell-division cycle as a simple model of organism development.

Comparative analysis of the CDC requires maintenance of populations of isogenic cells in culture. However, the yeast life-cycle has additional modes of reproduction which complicate such analysis. Members of *Saccharomyces* can develop as haploids or diploids (one or two copies of each chromosome) and can reproduce asexually through mitosis or sexually through meiosis. Mitosis results in the generation of a single daughter cell which buds off from the original mother cell. Meiosis is sporic and involves one round of diploid cell division followed by separation of the two diploid cells into four haploid cells (a tetrad of spores), which are then encased in an environmentally-resistant sac called an ascus. Yeast has two mating types, called a and α. Mating type is designated by different genes that can be located in and expressed from the mating type locus (MAT) found on chromosome III. Yeast must be diploid and heterozygous at the MAT locus in order to sporulate, and haploid yeast only has one or the other mating type. Thus haploids can arise either from cell division of an existing haploid or from meiosis of a diploid, while diploids can arise either from cell division of an existing diploid or conjugation of two haploid cells with opposite mating types. Moreover, wildtype haploids are able to switch mating types following mitosis (homothallism). In this way, populations founded by a single homothallic haploid cell have the potential to become sexual. This ability can be disrupted, however, through the deletion of the HO gene locus, which encodes a DNA endonuclease responsible for interchanging which genes are present at the MAT locus (*40*). Thus by deletion of the HO gene followed by forced sporulation to the haploid state, yeast populations can be maintained isogenically. In addition, haploid cells of one mating type are attracted to a peptide pheromone expressed by cells of the other mating type (a-factor and α-factor)

and respond by preparing for conjugation, which in part blocks cells from progressing beyond the G_1-phase (*41*). Thus artificial application of mating pheromone to a culture of isogenic, heterothallic yeast cells can be used to synchronize the majority of cells in the same CDC-phase within the mitotic cell cycle.

1.1.2 Population structure and evolutionary history of *Saccharomyces*

Outside of the laboratory, yeast cells are thought to propagate most often as diploids (*42, 43, 44*), typically by asexual mitosis (*45*). They also prefer to mate within-population (*46*). On its own, this behavior can lead to the formation of distinct populations which tend not to exchange genetic information. In addition, geographic isolation (*47*) of populations can contribute to the determination and reinforcement of such population structure in *Saccharomyces* as well as the species boundary between *S. cerevisiae* and *S. paradoxus*. Examples of populations evolving in this manner are commonly referred to as strains of yeast. Breweries and vineyards have been implicated in obvious examples of strain-specific evolution (*48*), but until recently the extent of reproductive isolation and the (natural and domestic) occurrence of yeast populations had been unclear, owing to a limited ability to classify these strains based on phenotypic differences. Instead of phenotypic markers, Naumov (*49, 50*) used reductions in the viability of hybrid offspring spores to define species boundaries within the *Saccharomyces sensu stricto* complex, in accordance with the standard biological species concept (*51*). Sniegowski *et al.* (*52*) isolated and genetically characterized dozens of yeast strains in both *S. cerevisiae* and *S. paradoxus* from two continents, demonstrating the existence of population structure within species. While *S. cerevisiae* has been isolated from a variety of ecological settings, *S. paradoxus* has only been found in woodland settings (*53*). A panmictic (free-mating) mode of reproduction along with geographic-associated reproductive isolation between North American and Eurasian iso-

lates demonstrates a continental population structure for *S. paradoxus*. In contrast, *S. cerevisiae* reproduces clonally and rarely mates outside its local population (low outcrossing rate) (*54, 55*), and distinct strains coexist at the same location (*56*).

Multilocus and whole-genome phylogenetic studies (*48, 55*) have established an evolutionary hierarchy of yeast populations, in which woodland *S. cerevisiae* isolates appear ancestral to most human-associated isolates and in general exhibit the most within-strain genetic variability. However, genetic characterization of woodland isolates from both species has revealed overall low levels of polymorphic variation among intronic sequences (*57*), suggesting few phenotypic differences and thus limited adaptive evolutionary potential. In contrast, sequencing of promoter regions of 8 transcription factor loci involved in CDC progression in woodland strains (Table 1.1) shows a 3.7-fold higher average pairwise nucleotide diversity in promoters ($\hat{\pi} = 7.08 \times 10^{-3}$) compared to introns ($\hat{\pi} = 1.89 \times 10^{-3}$). Woodland isolates also exhibit distinct mating preferences (*46*) and variability in cell size and CDC length (Figure 1.1), suggesting that this limited genetic variation is expressed phenotypically and may be implicated in the life history evolution of woodland populations. In addition, differences in thermal growth profiles have been found between species (*58*), and differential responses to freeze tolerance and copper sulfate exposure exist within species, between vineyard and woodland *S. cerevisiae* isolates (*59*), further suggesting the potential for adaptive evolution in yeast. Therefore, the genetic factors responsible for adaptive evolution may be found by comparative analysis of genomic loci across yeast populations.

1.1.3 Genome-scale molecular evolution in yeast

The genome of the laboratory strain of *S. cerevisiae* (S288c) (*25*) is composed of 16 nuclear chromosomes totaling 12.16 million nucleotides. This genome has 6607 annotated genes, 4819 of which have been verified as protein-coding, with 977 uncharacterized and

811 dubious open reading frames (ORFs) (*60*). The majority of genes encode single contiguous exons, but 316 genes contain single introns (*61*). Genome sequences have also been obtained for a few different yeast species, and their comparison has established an overall high degree of similarity among genomes in the *Saccharomyces sensu stricto* complex (*62*). Sequence similarity, simple gene structure, and relatively short regions of intergenic sequence all facilitate the identification of homologous nucleotide positions genome-wide across yeast genomes (*63*). Thus, initial investigations of genome-scale molecular evolution in yeast have focused on assessing levels of genetic divergence, refining annotations for gene boundaries, and identifying functional loci (*64*). Much of the variation found in yeast genomic DNA appears to be generated by point mutations, rather than by simple insertions, deletions, or larger structural alterations (*65*), or by gene flow across population or species boundaries (*54, 45*).

The ability to focus on identifying and counting alleles generated by point mutation at individual nucleotides among genomes further simplifies the assessment of evolutionary history among yeast genomes. The evolutionary forces which influence observed genetic variation can be determined by comparing the amount of variation at a locus to that expected solely due to random genetic drift following the neutral accumulation of mutations (*66*). This requires estimating the rate of introduction of new alleles into a population by mutation (mutation rate) as well as the number of generations separating the DNA sequences of interest (divergence time). Patterns of extreme variation in either direction provide evidence of evolution due to non-random forces, such as natural selection, as opposed to genetic drift (*67*). Alternatively, comparison of the average versus total nucleotide diversity at a locus can also reveal the effects of selection (*68*). However, the amount of genetic variation can also be influenced by non-selective factors, such as a population's size and mating preferences (population structure) (*67*), the degree of linkage disequilibrium in

a genome (genomic structure) (*69*), and developmental constraints (developmental structure) (*70*). Moreover, different genomic loci possess distinct functional roles in the cell, such that functional context affects the pattern of genetic variation (*71*). Proper assessment of evolutionary history should attempt to control for these factors.

Although comparisons of DNA sequences can be used in this way to infer yeast evolutionary history, it is very difficult, even for yeast, to predict how certain mutations have affected a cell's structure, function, molecular complement, dynamics, or ability to reproduce. In fact, the specific phenotypic effects of most DNA sequences are generally unknown, especially given the variety of possible environments to which a cell can be exposed (*72*) and the complexity of a cell's molecular interactions (*73*). Directly assessing the pattern of phenotypic variation complements the use of genetic variation in inferring evolutionary processes, especially in the context of potentially complex modes of evolution, such as the developmental expression of cellular phenotype.

1.1.4 Gene expression

The expression of cellular phenotype is the result of a complex network of molecular processes involved in decoding a cell's genotype via gene expression events and regulating the molecules produced. A gene expression event involves the execution of a series of regulatory processes, broadly grouped into the transcription of DNA into RNA and the translation of RNA into protein (*74*). Transcription initiates when transcriptional regulatory transfactors, together with an RNA polymerase complex, create physical associations near a gene's transcription start site, dependent on the sequence and location of transcription factor binding sequences of DNA, as well as chromatin structure. These associations help to polymerize a primary RNA transcript from a gene's DNA sequence. In yeast, the expression of genes transcribed into RNA transcripts is controlled predominantly by some 270

transcription factor proteins which bind to short, 5–20 nucleotide DNA motifs (*75*) found typically within 1000 nucleotides upstream ($5'$) of a gene's transcription start site (*76*). Thus, gene expression events form the basis of cellular (and molecular) phenotype.

In addition to exons, primary transcripts may include introns and untranslated regions (UTRs) that flank the $5'$ and $3'$ termini. Various post-transcriptional processes ($5'$-capping, poly-adenylation, intron splicing, degradation) control the maturation of this primary RNA, producing a mature messenger RNA transcript (mRNA). This transcript can then be exported from the cell nucleus, transported and spatially localized within the cell, or bound by ribosomes and translated into one or more proteins. Each protein is then subject to post-translational regulation, including protein modification, activation, localization, and degradation (*74*). All of these processes occur simultaneously within a cell, but they are also necessarily inter-dependent, temporally ordered, and influenced by environmental context. Thus, gene expression can be viewed as an integrated system of processes, which interact as a complex, dynamical system. I use the term epigenotype to refer to the specific architecture of gene expression event control encoded by a given cell's genotype.

1.1.5 Evolution of genome-wide gene expression

Many investigations of the evolution of genome-wide gene expression have focused on measuring and comparing RNA transcript levels (reflecting the number of RNA transcripts) present in a cell or tissue (reviewed in (*77*)), since all of the processes involved in a gene's expression are initiated by transcription and since measuring RNA levels genome-wide is technologically feasible. As with studies of genetic evolution, comparison of gene expression levels across related individuals or populations can provide insights into an organism's evolutionary history (*78*). However, gene expression is a quantitative trait, such that a sample of measurements represents a continuous distribution of expression levels, rather than a

small, discrete number of states. Thus a different approach is required to detect the effects of evolutionary forces on quantitative traits.

Two recent studies applied the idea of calibrating an observed distribution of expression levels against an expected distribution obtained after minimizing the influence of natural selection (*79*). Rifkin *et al.* (*80*) estimated expected distributions of gene expression levels using a set of replicate lines (mutation accumulation lines) of the fruit fly *Drosophila melanogaster*, where each line was evolved independently for 200 generations at a small population size (thereby minimizing the effects of natural selection). Comparison of a gene's observed expression variance among naturally-evolving flies with its expected variance among mutation accumulation lines provided a test for natural selection. They found that "although spontaneous mutations have the potential to generate abundant variation in gene expression, natural variation is relatively constrained." Denver *et al.* (*81*) performed a similar experiment using naturally-occurring species and experimental mutation-accumulation lines of the nematode worm *Caenorhabditis elegans*. They concluded that "strong stabilizing selection dominates the evolution of transcriptional change for thousands of *C. elegans* expressed sequences." Thus, negative or stabilizing natural selection appears to be the dominant mode of evolution operating on gene expression levels in these multicellular organisms, the effect of which is to limit the observed variation in the levels of gene expression across individuals.

Although these studies only recently verified the operation of stabilizing selection on genome-wide gene expression, the theory that stabilizing selection dominates phenotypic evolution in general was first promulgated in the mid-1900s by Waddington (*82*) and Schmalhausen (*83*), generally in the context of multicellular developmental processes such as morphogenesis. In contrast to the perspective that genes are autonomous entities, subject to independent evolutionary forces (*i.e.,* directional selection operating on mean expression

levels), they saw genes as parts that are coupled in pathways and networks which constitute developmental processes (*23, 84, 83*). Consideration of the epigenotype in this way led Waddington and Schmalhausen to suppose that stabilizing selection should dominate phenotypic evolution during organism development, because an evolutionary account of the diversity of species types requires that an organism recreate its developmental process reliably. Consequently, stabilizing selection could ensure stable expression of genes during development, while permitting changes during adult life-cycle phases by means of directional selection (*85*).

A gene's capacity to exhibit phenotypic variation (variability) can be assessed by observing its level of expression in different environments or genotypes (*17*). Limited expression level variation in a particular gene despite these perturbations reflects the epigenotype's ability to buffer against genetic or environmental changes (*86*). Holding genotype or environment constant, phenotypic variability can also be monitored throughout development, where changes in the magnitude or direction of variation as a function of time reveal an epigenotype's complexity due to the structure, composition, or dynamics of development (*70, 87*). Using a constant environment, Rifkin *et al.* (*80*) found differential variability in genome-wide gene expression at two developmental stages of the fly, concluding that developmental context affects gene expression evolution. In addition, analysis of primate species (*88*) revealed that although stabilizing selection appears to influence a large number of genes, evolution of gene expression for certain classes of genes, notably transcription factors, appears more variable in humans compared to other primates. These observations argue that the effect of stabilizing selection depends on an organism's position in the life-cycle and suggest that changes in the regulation of gene expression play an important role in gene expression evolution, as expected if the theories of Waddington and Schmalhausen are correct.

There are actually two distinct but related modes of negative selection that operate on phenotypic variation: stabilizing selection, which directly limits a population's expressed variation in a trait; and canalizing selection or canalization, which indirectly limits expressed variation by constraining the underlying developmental process which generates the possible trait variants (*89, 17*). While both modes result in decreased phenotypic variation, canalization achieves this by altering a trait's capacity to exhibit variation (variability). Distinguishing between these modes has proven to be a difficult task, since simply comparing levels of natural variation to those expected without the influence of selection is not sufficient to distinguish whether the underlying phenotypic variability has changed (*17*). Nevertheless, examples of canalization have been identified experimentally (*82, 90*), and more recently canalization has been implicated in the evolution of genome-wide gene expression (*80*).

The evolution of genome-wide gene expression in natural yeast populations has not yet been characterized, but these previous studies predict that yeast gene expression may evolve by stabilizing selection, as in other multicellular organisms. Moreover, since the yeast cell cycle is developmentally robust to environmental and genetic perturbations (*72, 91, 35*) and serves a critical role as the most fundamental developmental process for the construction of multicellular organisms, temporal patterns of genome-wide gene expression may also exhibit signatures of canalization.

1.1.6 Life-cycle evolution and heterochrony

Waddington's perspective that "changes in genotypes only have ostensible effects in evolution if they bring with them alterations in the epigenetic processes by which phenotypes come into being" (*84*) holds genome evolution in a broader context of the evolution of ontogeny, the developmental progression of an individual's life-cycle. In this regard Gavin de

Beer's concept of heterochrony (*92*) becomes relevant. de Beer argued that heterochrony, or change in the relative timing between developmental processes, serves as the dominant mode of life-cycle evolution, and he developed a framework to classify the effects of different kinds of heterochrony (*4*). Seeing development as a composition of inherently modular processes, de Beer argued that ultimately there exist only two sources of evolutionary change in ontogeny, the introduction of a novel process or the alteration of an existing process, and that they can occur within any developmental module. Drawing from Goldschmidt's work, he proposed a mechanism for heterochrony involving changes in the expression of genes which determine the rates of developmental processes (*93*). Although both the biological mechanisms of gene expression and de Beer's heterochrony classification scheme have been substantially revised in recent decades (*15, 94, 95, 96*), heterochronic evolution of molecular gene expression levels across related species has been observed and associated with life-cycle evolution (*97*). Changes in the timing of gene expression during development (brought about for example by genetic mutation) may thus contribute to evolutionary change, offering a possible explanation for broad changes in gene expression associated with adaptation among highly related populations, as have been shown in yeast (*98*).

1.2 Overview of Dissertation

Genome-wide regulation of gene expression involves a dynamic epigenetic structure which generates an organism's life-cycle. Although changes in gene expression during development have broad effects on many basic phenomena including cell growth and differentiation (*34, 99*), morphogenesis (*100*), and disease progression (*101*), the evolutionary forces influencing gene expression dynamics and gene regulation remain largely unknown, due

to the nature of gene expression as a complex, quantitative trait (*102*). Moreover, since gene expression is regulated differentially over time, the effects of evolutionary forces may be influenced by developmental context. To advance the understanding of evolution in the context of the life-cycle, this dissertation presents two experimental investigations of the evolution of genes and related structural regions and time-dependent gene expression, using the budding yeasts *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* and their mitotic cell-division cycles as model organism and life-cycle. Comparative methodologies are employed to analyze genome-wide patterns of genetic and phenotypic diversity within and between species.

Chapter 2 of this dissertation describes the application of population genetics methodology to recent genome sequencing efforts, in order to investigate the evolutionary history of the yeast genome. Although systematic population genetic accounts of yeast evolutionary history have been made (*69, 48, 47, 45*), they preceded the availability of entire genome sequences and instead focused on relatively small, multilocus samples of the genome. Recently, the complete genome sequences of several dozen isolates of two species of budding yeast became available (*55*), providing an opportunity to study the evolutionary history of the entire yeast genome within and between two closely related species. With Junhyong Kim, I identified the protein-coding genes within each genome sequence and subdivided each gene into a collection of genic structural regions. Distinguishing among structural regions allowed us to characterize the evolutionary history of different regions within each gene and differentiate between modes of evolution constrained by a region's functional context.

In chapter 3, the focus shifts from the evolution of the DNA-based encoding of genes to evolution of their expression. Recent studies have verified the theory of stabilizing selection on gene expression at the transcriptional level (*103, 80, 81*), while suggesting that

the strength of stabilizing selection on gene expression levels may be influenced by an organism's developmental stage. Although various studies have used yeast to investigate the effects of environmental (*72*), genetic (*104*), and developmental (*105, 35*) factors on gene expression, the evolution of gene expression among natural yeast populations (*106*), and the properties of the epigenotype which influence gene expression evolution (*107, 108*), no study has considered whether genome-wide gene expression exhibits time-dependent signatures of evolutionary variability, which could demonstrate clearly whether natural selection is influenced by organism development. In addition, no study has yet characterized natural genome-wide gene expression variability in yeast (but see (*107*) for a study of mutational variation). With Chantal Francis, Paul Sniegowski, and Junhyong Kim, I measured natural variation in mRNA gene expression as a genome-wide time-series through the mitotic cell-division cycle of nine closely related woodland lines of the budding yeast *S. cerevisiae* and one outgroup *S. paradoxus*. We analyzed this multi-genome time-series data set using comparative genomics approaches to characterize time-dependent signatures of evolutionary gene expression variation, elucidate the modes of evolution affecting gene expression, and propose a model for the time-dependent architecture of the yeast epigenotype. In analyzing these time-series data we viewed the yeast cell-division cycle as a simple developmental and complex dynamical system, with regular schedules of molecular execution (*32, 109*), robustness to perturbations (*72, 91, 35*), and well-defined end goal of cell replication and division.

Chapter 4 of this dissertation provides a brief summary of all research findings, comments about methodological limitations, and possible future directions of research that stem from this work.

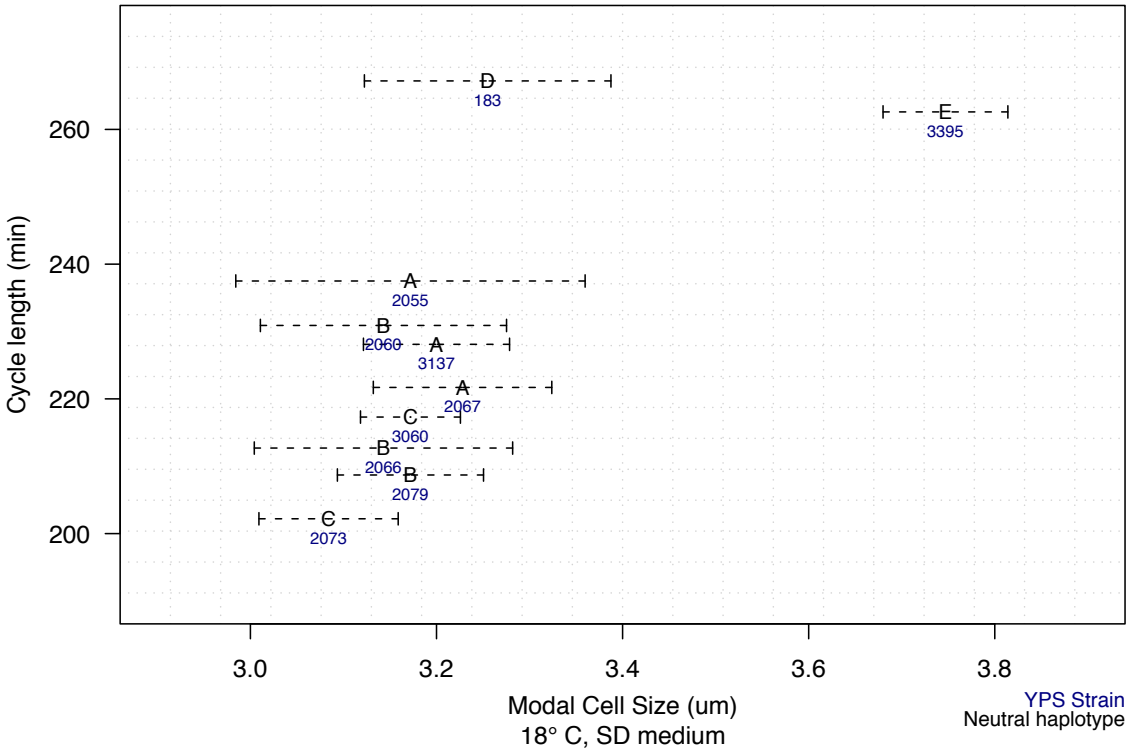


Figure 1.1: Variability in CDC length and cell size in late G_1-phase for 10 *Saccharomyces* isolates

The 10 yeast isolates are MATa haploid and include 8 woodland isolates and 1 laboratory strain (183) of *S. cerevisiae* and 1 woodland isolate of *S. paradoxus* (3395) (see Table 3.1). Isolates are identified by name and by haplotype group. Haplotype groups A, B, and C denote woodland *S. cerevisiae* isolates (*57*). Measurements were made on cultures of each isolate grown in 18° C minimal medium (SD). Late G_1-phase cell size was estimated during α-factor arrest using a Coulter counter. There are 5 biological replicate measurements of cell size for each isolate culture. Cell size varies significantly among the 8 woodland *S. cerevisiae* isolates (ANOVA, $P < 0.05$). CDC length was estimated using a damped sine-wave regression on the budding index of each isolate, following α-factor arrest and release. CDC length estimates for *S. cerevisiae* isolates were obtained by F. Ge.

Table 1.1: Genetic variability in proximal promoters of transcription factor loci among woodland *S. cerevisiae* isolates

Locus	Length	S	S/length ($\times 10^{-3}$)	π ($\times 10^{-3}$)
Far1 5′	720	2.0	8.61	3.84
Far1 3′	600	7.0	14.63	8.06
Mbp1 5′	660	1.0	4.64	3.73
Ifh1 5′	822	10.0	14.83	7.60
Whi5 5′	406	6.0	19.66	12.90
Whi3 5′	546	1.0	1.83	0.98
Sfp1 5′	540	10.0	18.52	9.52
Sfp1 3′	375	7.0	18.67	9.98
Average	584	5.5	12.67	7.08

The DNA sequence of each locus was obtained for 8 haploid woodland *S. cerevisiae* isolates (see Table 3.1). Sequences correspond to the proximal (flanking) $5'$ or $3'$ noncoding region adjacent to a protein-coding locus, as indicated. Comparison of 3 *S. cerevisiae* genomes (S288c, RM11-1a, YJM789) indicated that these loci harbor elevated levels of genetic variation, suggesting that they are also polymorphic among woodland populations. Length reports the number of nucleotides sequenced. S is the total number of segregating sites within a locus. S/length is the per-nucleotide fraction of segregating sites, reported per kilobase of sequence. π is the average per-nucleotide pairwise sequence diversity, reported per kilobase of sequence. π values were computed by aligning all pairs of sequences using global alignment, counting the number of identical nucleotides for each pair, averaging over all pairs, and dividing by the sequence length.

References

1. von Linné, C. *Systema Naturae per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, cum Charateribus, Differentiis, Synonymis, Locis*. Holmiae: Impensis Direct, (1758).

2. Haeckel, E. H. P. A. *Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*. G. Reimer, Berlin, (1866).

3. Baer, K. E. v. *Über Entwickelungsgeschichte der Thiere; Beobachtung und Reflexion*. Bei den Gebrüdern Bornträger, Königsberg, (1837).

4. De Beer, G. *Embryology and Evolution*. The Clarendon Press, Oxford, (1930).

5. Kimura, M. *Nature* **217**(5129), 624–626 (1968).

6. Frazer, K. A., Ballinger, D. G., Cox, D. R., et al. *Nature* **449**(7164), 851–861 (2007).

7. Lande, R. *Evolution* **30**(2), 314–334 (1976).

8. Gould, S. J. *Paleobiology* **6**(1), 119–130 (1980).

9. Lynch, M. *Evolution* **40**(5), 915–935 (1986).

10. Hall, B. K., Pearson, R. D., and Müller, G. *Environment, Development, and Evolution: toward a Synthesis*. MIT Press, Cambridge, MA, (2004).

11. Hooke, R. *Micrographia: or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries thereupon*. John Martyn and James Allestry, November (1664).

12. Bonner, J. T. *Size and Cycle: An Essay on the Structure of Biology*. Princeton University Press, (1965).

13. Stearns, S. C. *The Evolution of Life Histories*. Oxford University Press, New York, NY, (2004).

14. Laubichler, M. D. and Maienschein, J., editors. *From Embryology to Evo-Devo: A History of Developmental Evolution*. The MIT Press, Cambridge, MA, (2007).

15. Gould, S. J. *Ontogeny and Phylogeny*. Harvard University Press, Cambridge, MA, (1977).

16. Raff, R. A. *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. University of Chicago Press, Chicago, IL, (1996).

17. Willmore, K. E., Young, N. M., and Richtsmeier, J. T. *Evolutionary Biology* **34**, 99–120 (2007).

18. Gerhart, J. and Kirschner, M. *Cells, Embryos, and Evolution: toward a Cellular and Developmental Understanding of Phenotypic Variation and Evolutionary Adaptability*. Blackwell Science, Malden, MA, (1997).

19. Oyama, S., Griffiths, P. E., and Gray, R. D. *Cycles of Contingency: Developmental Systems and Evolution*. MIT Press, Cambridge, MA, (2001).

20. Waddington, C. H. *The Evolution of an Evolutionist*. Cornell University Press, Ithaca, NY, (1975).

21. Goldschmidt, R. *The Scientific Monthly* **46**(3), 268–273 (1938).

22. Goldschmidt, R. *The Material Basis of Evolution*. Yale University Press, New Haven, CT, (1982).

23. Waddington, C. H. *Nature* **147**, 108–110 (1941).

24. Ringo, J. *Fundamental Genetics*. Cambridge University Press, New York, (2004).

25. Goffeau, A., Barrell, B. G., Bussey, H., et al. *Science* **274**(5287), 546, 563–7 (1996).

26. Mortimer, R. K. and Hawthorne, D. C. *Annu. Rev. Microbiol.* **20**, 151–168 October (1966).

27. Hartwell, L. H., Culotti, J., Pringle, J. R., and Reid, B. J. *Science* **183**(4120), 46–51 (1974).

28. Giaever, G., Chu, A. M., Ni, L., et al. *Nature* **418**(6896), 387–391 (2002).

29. Dujon, B., Sherman, D., Fischer, G., et al. *Nature* **430**(6995), 35–44 (2004).

30. Replansky, T., Koufopanou, V., Greig, D., and Bell, G. *Trends Ecol Evol* **23**(9), 494–501 (2008).

31. Landry, C. R., Townsend, J. P., Hartl, D. L., and Cavalieri, D. *Mol Ecol* **15**(3), 575–591 (2006).

32. Hartwell, L. H. and Weinert, T. A. *Science* **246**(4930), 629–634 (1989).

33. Gilbert, S. F. *Developmental Biology*. Sinauer Associates, Inc., 6th edition, (2000).

34. Jorgensen, P. and Tyers, M. *Curr Biol* **13**, 1014–1027 December (2004).

35. Orlando, D. A., Lin, C. Y., Bernard, A., et al. *Nature* **453**(7197), 944–947 (2008).

36. Soranzo, N., Zampieri, M., Farina, L., and Altafini, C. *BMC Syst Biol* **3**(18) February (2009).

37. Hamatani, T., Carter, M. G., Sharov, A. A., and Ko, M. S. H. *Dev Cell* **6**(1), 117–131 (2004).

38. Lopez-Maury, L., Marguerat, S., and Bahler, J. *Nat Rev Genet* **9**(8), 583–593 August (2008).

39. Sangster, T. A., Lindquist, S., and Queitsch, C. *Bioessays* **26**(4), 348–362 (2004).

40. Herskowitz, I. *Microbiol Rev* **52**(4), 536–553 (1988).

41. Chan, R. K. and Otte, C. A. *Mol Cell Biol* **2**(1), 21–29 (1982).

42. Johnson, L. J., Koufopanou, V., Goddard, M. R., Hetherington, R., Schäfer, S. M., and Burt, A. *Genetics* **166**, 43–52 January (2004).

43. Gerstein, A. C., Chun, H.-J., Grant, A., and Otto, S. P. *PLoS Genet* **2**(9) (2006).

44. Mortimer, R. K., Romano, P., Suzzi, G., and Polsinelli, M. *Yeast* **10**(12), 1543–1552 (1994).

45. Tsai, I. J., Bensasson, D., Burt, A., and Koufopanou, V. *Proc Natl Acad Sci U S A* **105**(12), 4957–4962 (2008).

46. Murphy, H. A., Kuehne, H. A., Francis, C. A., and Sniegowski, P. D. *Biol Lett* **2**(4), 553–556 (2006).

47. Kuehne, H. A., Murphy, H. A., Francis, C. A., and Sniegowski, P. D. *Curr Biol* **17**(5), 407–411 (2007).

48. Fay, J. C. and Benavides, J. A. *PLoS Genet* **1**(1), 66–71 (2005).

49. Naumov, G. *Journal of Industrial Microbiology* **17**, 295–302 (1996).

50. Fischer, G., James, S. A., Roberts, I. N., Oliver, S. G., and Louis, E. J. *Nature* **405**(6785), 451–454 (2000).

51. Mayr, E. *Systematics and the Origin of Species*. Columbia University Press, New York, (1942).

52. Sniegowski, P. D., Dombrowski, P. G., and Fingerman, E. *FEMS Yeast Res* **1**(4), 299–306 (2002).

53. Naumov, G. I., Naumova, E. S., and Sniegowski, P. D. *Can J Microbiol* **44**, 1045–1050 (1998).

54. Ruderfer, D. M., Pratt, S. C., Seidel, H. S., and Kruglyak, L. *Nat Genet* **38**(9), 1077–1081 (2006).

55. Liti, G., Carter, D. M., Moses, A. M., et al. *Nature* **458**(7236), 337–341 (2009).

56. Koufopanou, V., Hughes, J., Bell, G., and Burt, A. *Philos Trans R Soc Lond B Biol Sci* **361**(1475), 1941–1946 (2006).

57. Kuehne, H. A. *The Genetic Structure and Biogeography of Natural Saccharomyces Populations*. PhD thesis, University of Pennsylvania, (2005).

58. Sweeney, J. Y., Kuehne, H. A., and Sniegowski, P. D. *FEMS Yeast Res* **4**(4-5), 521–525 (2004).

59. Fay, J. C., McCullough, H. L., Sniegowski, P. D., and Eisen, M. B. *Genome Biol* **5**(4), R26 (2004).

60. *Saccharomyces* Genome Database. January (2008).

61. Lopez, P. J. and Seraphin, B. *Nucleic Acids Res* **28**(1), 85–86 (2000).

62. Kellis, M., Birren, B. W., and Lander, E. S. *Nature* **428**(6983), 617–624 (2004).

63. Conant, G. C. and Wolfe, K. H. *Genetics* **179**(3), 1681–1692 (2008).

64. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. *Nature* **423** May (2003).

65. Lynch, M., Sung, W., Morris, K., et al. *Proc Natl Acad Sci U S A* **105**(27), 9272–9277 (2008).

66. Li, W.-H. *Molecular Evolution*. Sinauer Associates, Sunderland, MA, (1997).

67. Gillespie, J. H. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, second edition, (2004).

68. Tajima, F. *Genetics* **123**, 585–595 November (1989).

69. Fay, J. C. and Benavides, J. A. *Genetics* **170** August (2005).

70. Smith, J. M., Burian, R., Kauffman, S., et al. *The Quarterly Review of Biology* **60**(3), 265–287 September (1985).

71. Cliften, P., Sudarsanam, P., Desikan, A., et al. *Science* **301**(5629), 71–76 July (2003).

72. Gasch, A. P., Spellman, P. T., Kao, C. M., et al. *Mol Biol Cell* **11**(12), 4241–4257 (2000).

73. Goll, J. and Uetz, P. *Genome Biol* **7**(6), 223 (2006).

74. Lewin, B. *Genes VII*. Oxford University Press, Oxford, (2000).

75. MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. *BMC Bioinformatics* **7**, 113 (2006).

76. Beer, M. A. and Tavazoie, S. *Cell* **117**(2), 185–198 (2004).

77. Gilad, Y., Oshlack, A., and Rifkin, S. A. *Trends Genet* **22**(8), 456–461 (2006).

78. Oleksiak, M. F., Churchill, G. A., and Crawford, D. L. *Nat Genet* **32**(2), 261–266 (2002).

79. Lynch, M. and Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA, (1998).

80. Rifkin, S. A., Houle, D., Kim, J., and White, K. P. *Nature* **438**(7065), 220–223 (2005).

81. Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M., and Thomas, W. K. *Nat Genet* **37**(5), 544–548 (2005).

82. Waddington, C. H. *Nature* **150**, 563–565 (1942).

83. Schmalhausen, I. I. *Factors of Evolution: the Theory of Stabilizing Selection*. The Blakiston Company, Philadelphia, PA, (1949).

84. Brown, R. and Danielli, J. F., editors. *Symposia Of The Society For Experimental Biology - Number VII Evolution*. Academic Press, New York, (1953).

85. Gilbert, S. F. *American Zoologist* **40**(5), 729–737 November (2000).

86. Schwenk, K. and Wagner, G. P. *American Zoologist* **41**, 552–563 (2001).

87. Brakefield, P. M. *Trends Ecol Evol* **21**(7), 362–368 (2006).

88. Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P., and White, K. P. *Nature* **440**(7081), 242–245 (2006).

89. Gibson, G. and Wagner, G. *Bioessays* **22**, 372–380 (2000).

90. Rutherford, S. L. and Lindquist, S. *Nature* **396**(6709), 336–342 (1998).

91. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., and others (52 co authors). *Science* **285**(5429), 901–906 (1999).

92. Rice, S. H. *Encyclopedia of Evolution*, chapter Heterochrony, http://www.oxford–evolution.com/entry?entry=t169.e191. Oxford University Press, University of Pennsylvania, e-reference edition edition (2005).

93. Brigandt, I. *J Exp Zoolog B Mol Dev Evol* **306**(4), 317–328 (2006).

94. Alberch, P., Gould, S. J., Oster, G. F., and Wake, D. B. *Paleobiology* **5**(3), 296–317 (1979).

95. Slatkin, M. *Evolution* **41**(4), 799–811 (1987).

96. Raff, R. A. and Wray, G. A. *J Evol Biol* **2**, 409–434 (1989).

97. Kim, J., Kerr, J. Q., and Min, G. S. *Proc Natl Acad Sci U S A* **97**(1), 212–216 (2000).

98. Ferea, T. L., Botstein, D., Brown, P. O., and Rosenzweig, R. F. *Proc Natl Acad Sci U S A* **96**(17), 9721–9726 (1999).

99. Boheler, K. *J Cell Physiol* **221**(1), 10–17 (2009).

100. Davidson, E. H. and Erwin, D. H. *Science* **311**, 796–800 10 February (2006).

101. Ross, D. T., Scherf, U., Eisen, M. B., et al. *Nat Genet* **24**(3), 227–235 (2000).

102. Rockman, M. V. and Kruglyak, L. *Nat Rev Genet* **7**(11), 862–872 (2006).

103. Rifkin, S. A., Kim, J., and White, K. P. *Nat Genet* **33**(2), 138–144 (2003).

104. Hughes, T. R., Marton, M. J., Jones, A. R., et al. *Cell* **102**(1), 109–126 (2000).

105. Rifkin, S. A., Atteson, K., and Kim, J. *Funct Integr Genomics* **1**(3), 174–185 (2000).

106. Townsend, J. P., Cavalieri, D., and Hartl, D. L. *Mol Biol Evol* **20**(6), 955–963 (2003).

107. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J., and Hartl, D. L. *Science* **317**(5834), 118–121 (2007).

108. Tirosh, I., Reikhav, S., Levy, A. A., and Barkai, N. *Science* **324**(5927), 659–662 (2009).

109. Klevecz, R. R., Li, C. M., Marcus, I., and Frankel, P. H. *FEBS J* **275**(10), 2372–2384 (2008).

Chapter 2

Evolutionary Dynamics of the *Saccharomyces* Genome and a Resource for Population Genomics

2.1 Abstract

Background

The evolutionary dynamics of the *Saccharomyces sensu stricto* clade is understood incompletely. With the release of dozens of yeast genome sequences, it has become possible to perform a systematic study of the micro and macro-evolutionary dynamics of yeast gene structure, which should provide broad insights into the interface between micro-evolution and speciation.

Methodology/Principal Findings

We present a genome-wide analysis of yeast evolutionary dynamics with estimates of polymorphism and divergence based on sequence alignments for 10 categories corresponding to unique genic structural regions for 6,575 *S. cerevisiae* and 5,250 *S. paradoxus* protein-coding genes from 67 genome sequence assemblies. We find an overall signature of moderate to strong purifying selection within all structural regions of both species, which falls strongest on coding sequence and proximal non-transcribed regions (e.g. promoters) and

weakest on the intron and untranslated regions. The differential response of selection to similar patterns of variation in non-coding proximal and untranslated regions indicates the influence of genic structural regions on evolutionary mode. Interspecies divergence is limited overall but elevated in proximal regions, despite evidence of strong purifying selection there. Although few Gene Ontology terms associate with regions exhibiting excess divergence, the 3′ untranslated regions of ribosomal proteins show excess divergence in combination with reduced promoter divergence. Using a subset of neutrally evolving intronic sequence, divergence time between the two species is an estimated 250,000 years, with current population sizes of 16 and 42 million for *S. cerevisiae* and *S. paradoxus*. To foster future comparative and phenotype association studies in yeast, all relevant DNA sequences, multiple alignments, and statistics are provided through the Budding Yeast Gene Evolution database (http://yeastpopgenomics.org).

Conclusions/Significance

The genomes in *Saccharomyces* populations are evolving under moderate purifying selection within species and limited divergence between species. Despite this limitation, alterations in transcriptional regulation appear to have contributed the most to sequence divergence between species, while differences in the post-transcriptional regulation of ribosomal genes may have altered the timing of each species' life-history developmental transition from vegetative growth to reproduction.

2.2 Introduction

Despite the historical importance of budding yeast in science, industry, and culture, the natural history of the *Saccharomyces sensu stricto* clade has received little attention until the

past decade. Hundreds of *Saccharomyces* strains have now been isolated from several environments, including vineyards, fruit, saké fermenters, human hosts, and oak trees (*1, 2, 3*). Previous studies found that *S. cerevisiae* populations appear to segregate by ecotype, within which levels of genetic diversity vary. Woodland populations generally retain the most variability and vineyard and saké populations the least (*4*). Woodland isolates of *S. cerevisiae* and *S. paradoxus* show a high level of inbreeding, but *S. cerevisiae* appears to reproduce clonally while *S. paradoxus* exhibits sexual reproduction on a continental scale (*5*). Consequently diversity within geographically similar woodland populations of both species is low, but has been shown to increase with distance in *S. paradoxus* (*6, 7*). Moderate levels of interlocus recombination were found among Pennsylvania oak and vineyard *S. cerevisiae* isolates and English *S. paradoxus* isolates (*8, 9*). Both sympatric and allopatric isolates of *S. paradoxus* yield reduced hybrid ascospore viability, indicative of postzygotic reproductive isolation within and between continents (*10, 6*). At the same time mating dynamics between species differ, suggesting prezygotic barriers exist as well (*11*).

While these studies analyzed multilocus sequence data from a variety of strains, they were nevertheless performed using a small subset of the entire yeast genome. With the recent release of genome sequences for several dozen *S. cerevisiae* and *S. paradoxus* strains (*12*), it has become possible to investigate both population and gene structure at a whole-genome scale. While protein-coding DNA naturally encodes polypeptides, various gene-associated non-coding sequences are functionally important for the proper production and maintenance of these proteins. These promoters, untranslated regions (UTR), and introns tend to exhibit unique regulatory roles, so understanding how they are shaped by evolutionary forces should provide broad insights into the interface between micro-evolution and speciation (*13*).

Our results present a genome-wide analysis of the evolutionary dynamics of *Saccha-*

romyces and provide estimates of polymorphism and divergence based on sequence alignments for 10 categories corresponding to unique genic structural regions (GSRs) for 6,575 *S. cerevisiae* and 5,250 *S. paradoxus* protein-coding genes from 39 *S. cerevisiae* and 28 *S. paradoxus* publicly available genome sequence assemblies. We analyze patterns of polymorphism and divergence to infer natural selection in the context of these GSR categories, and associate this variation with gene ontology (GO) terms. We discuss the relationship between evolutionary forces and population structure, and suggest two mechanisms for species divergence. We also estimate the divergence time between species and population coalescence times within species. To foster future comparative and phenotype association studies in yeast, the Budding Yeast Gene Evolution database (http://yeastpopgenomics.org) has been created to provide access to all relevant DNA sequences, multiple alignments, and statistics.

2.3 Results

2.3.1 Identification and partitioning of protein-coding loci

We identified an average of 4,236.8 and 3,379.2 open reading frames (ORF) from each of 39 *S. cerevisiae* and 28 *S. paradoxus* publicly available genome sequence assemblies (see Materials and Methods). The original sequences were generated using a whole-genome shotgun approach, and the majority of sequences have approximately $1\times$ coverage (see Figure 2.2). After combining sequences from all genomes, we obtained population samples for 6,575 *S. cerevisiae* and 5,250 homologous *S. paradoxus* ORFs, comprising 97.9% and 78.1% of the 6719 ORFs identified in *S. cerevisiae*, respectively. Although this includes verified and putative ORFs, as well as transposable (Ty) elements, a Fisher's Exact test indicated that a significant percentage (79.6%) of the identified *S. paradoxus* ORFs

corresponds to verified ORFs in *S. cerevisiae* ($P < 10^{-230}$).

88 Ty elements were recovered from *S. cerevisiae* genomes, one of which was found in *S. paradoxus* (YHL009W-A, a TyA Gag gene). Although Ty1, Ty3, and Ty5 elements have been identified in limited numbers in *S. paradoxus*, their presence among *S. paradoxus* isolates is much more variable than among *S. cerevisiae* isolates (*14*). Since our gene identification algorithm uses a single reference genome to identify ORFs in unannotated genome assemblies for each species, it is likely the *S. paradoxus* reference we used (NRRL Y-17217, the only annotated *S. paradoxus* genome) lacks many of the Ty elements found in other *S. paradoxus* strains.

Loci associated with each ORF were then partitioned into 8 primary genic structural regions (GSRs): coding sequence (CDS), intron, untranslated regions (5′/3′ UTR), proximal non-coding sequence ranging from UTR terminals to 500 bases up/downstream (5′/3′ proximal), and distal non-coding sequence from 500 bases to the next adjacent CDS start or stop codon (5′/3′ distal). We also included 2 composite categories which encompass all contiguous non-coding sequences between two ORFs (5′/3′ composite). We refer to the labels denoting each GSR (CDS, intron, UTR, proximal, distal, composite) as GSR categories.

Since non-coding GSRs may contain elements functional for multiple neighboring ORFs, no effort was made to associate a non-coding sequence exclusively to a single gene. In this way a DNA sequence may be associated with different GSR categories from two adjacent ORFs (e.g. the 5′ distal of one gene may also serve as the 3′ UTR of another). Based on average GSR lengths and diversity estimates, however, each GSR at every locus is expected to contain predominantly unique DNA sequence. GO annotation of DNA sequences is handled in a similar manner. Just as an individual DNA sequence may be associated with two GSR categories, the same sequence may be annotated as one GO term with respect to

one ORF and a second term with respect to an adjacent ORF.

Table 2.6 lists the number of sequences obtained for each GSR category for both species. 55.7% of *S. cerevisiae* introns were not identified in *S. paradoxus*. Presumably these introns are present but, due to changes in gene structure or sequence divergence, could not be aligned with *S. cerevisiae* introns and therefore could not be included in a comparative analysis. Overall this data set yields an average of 21.6 sequence samples (taxa) for 10 GSR categories of nearly every gene in the *Saccharomyces* genome.

2.3.2 Heterozygosity within GSR categories

The first step in evaluating the evolutionary history of a genetic locus is to estimate the level of nucleotide variation, which is often described by the site heterozygosity parameter $\theta = 4N\mu$ (*15*). Since θ is proportional to the product of population size (N) and spontaneous per-nucleotide mutation rate (μ), differences in θ between populations may indicate differing population sizes, mutation rates, or selection pressures (*16*). We estimated θ in two ways, using the average number of pairwise differences ($\hat{\pi}$) and the scaled number of segregating sites ($\hat{\theta} = S/a1$), per nucleotide site (see Materials and Methods).

The genome-wide levels of polymorphism within GSR categories are shown in Figure 2.1; statistics are presented in Table 2.6. *S. cerevisiae* exhibits a genome-wide average $\hat{\theta}$ equal to 0.0465, with an average of 20.5 differences per kilobase (kb) between a pair of homologous sequences, whereas *S. paradoxus* has an average $\hat{\theta}$ of 0.0439 and an average of 28.6 pairwise differences per kb. This corresponds to genome-wide average pairwise differences of 2.1% and 2.9% within species, while divergence between species is 22.1% (11% in coding sequence and 26.7% in non-coding). Significant differences in both estimates of nucleotide variation were found between species for all GSR categories, with the exception of the 3′ UTR (using $\hat{\theta}$) and distal non-coding GSR categories (using $\hat{\pi}$) (t-test,

$P < 0.01/10$). These differences are generally small, especially for CDS, but larger shifts can be seen for 5′ and 3′ proximal and distal GSR categories, indicating the potential for differential selection pressure between species.

S. cerevisiae has relatively more segregating sites, whereas *S. paradoxus* has more pairwise variation. This relative ordering is seen for all GSR categories except the 5′ UTR, 3′ UTR, and intron (see Table 2.1), where *S. paradoxus* shows greater variation in $\hat{\theta}$ (with differences of 0.0060, 0.0026, and 0.0140, respectively), suggesting a deficit of deleterious mutations in these GSR categories in *S. cerevisiae*. The fact that this difference is more than twice as high in introns, and the fact that introns have the highest $\hat{\pi}$ within *S. paradoxus* are both consistent with the presumed selective neutrality of introns (*17*). These results could be biased by sampling differences between species, as the majority of *S. paradoxus* isolates derive from oak trees, whereas *S. cerevisiae* isolates come from a diverse range of environments. Nevertheless *S. paradoxus* displays more pairwise variation ($\hat{\pi}$), which is much less sensitive to short-term fluctuations in allele frequency caused by deleterious mutations (*18*). These results support the claim of population structure differences between species (*5*) and suggest that deleterious mutations may exist in greater abundance in *S. cerevisiae*.

2.3.3 Evolution of GSR categories within species

To determine whether observed polymorphism is compatible with neutral evolution, we computed Tajima's D statistic (*18*) for every GSR in each category per species. This statistic takes advantage of discrepancy between $\hat{\theta}$ and $\hat{\pi}$ in estimating θ to test the neutral mutation hypothesis. Although no GSR categories (average of genome-wide D values) show significant departures from neutrality after a 5% false discovery rate correction per species, CDS, 5′ proximal, and 3′ proximal categories of *S. cerevisiae* exhibit the most extreme values ($P < 0.05$), in the direction suggestive of purifying selection (see below

and Table 2.6). A significant correlation of 0.62 between Tajima's D and $\hat{\pi}$ ($P < 0.01$) is seen across GSR categories, which may reflect strong linkage across genic structural regions (*19, 20*). This relationship holds much better for *S. paradoxus* ($R = 0.64$) than for *S. cerevisiae* ($R = 0.17$) however. The lack of strong correlation in *S. cerevisiae* may reflect relatively compact regions of linkage (notably in vineyard strains) (*22*), but in that case this correlation should be more pronounced in *S. paradoxus*, which has a predominantly sexual reproductive mode (*5*) and little apparent population structure (*12*). Instead we observe a strong correlation between D and $\hat{\pi}$ in *S. paradoxus*, which suggests that species-specific differences in demographics (*e.g.* expanding population size or mating preferences) or genome structure (*e.g.* linkage or recombination rate) may influence effects of selection on each GSR category (*21*).

Evidence for genome-wide purifying selection

Given that the assumptions of Tajima's D test are met (populations are constant-sized and randomly mating with no selection, no recombination, and loci have infinite sites), two observations suggest that the observed negative Tajima's D values indicate the overall presence of purifying rather than positive selection. First, both species show comparable, moderately negative D values (averaging -1.58 overall), which argues against either a split between purifying selection in one species and positive selection in the other or an overall signature of positive selection for both species. Moreover, a slight excess of low-frequency alleles within a population is expected by the nearly-neutral theory, which posits that most alleles carry a neutral or slightly deleterious fitness cost (*23*).

To test whether the direction of selection is purifying, we looked at correlations between Tajima's D values and interspecies sequence divergence. Since positive selection decreases the frequency of a position's ancestral allele, ultimately replacing it with a different allele,

an allele under positive selection should also diverge from the most frequent allele at the same locus in a closely related species, assuming it is not also under positive selection. Thus, a positive correlation is expected between Tajima's D and sequence divergence for GSR categories under positive selection, while no correlation or a negative correlation is expected for categories under negative selection. Correlations were computed by GSR category using sequences with significantly negative D values. The rate of substitution ρ between *S. cerevisiae* and *S. paradoxus* was used as the divergence metric; estimates were obtained by global alignment of one randomly selected sequence from each species (see Table 2.7). In this way we expect to align the major alleles of both species (*24*). Individual and pooled GSR category correlations are weak (pooled $R = -0.011,\ P < 0.22$). CDS and 5′ proximal showed the largest correlations of 0.117 and 0.100, while intron showed the lowest correlation with -0.133. The pooled correlation is also smaller than that shown using sequences lacking significance in Tajima's D (pooled $R = 0.058,\ P < 10^{-7}$). Given the overall negative relationship between D and ρ, we conclude that in general, negative values of Tajima's D indicate a departure from neutrality towards purifying selection. The possibilities that these negative values instead identify linked selected sites or result from an expanding population size are evaluated in the Discussion.

Evidence for genic structural regions under balancing selection

Although we find no evidence that GSR categories evolve under balancing selection, several individual genes (1.9% overall) do exhibit an excess of high-frequency variation, notably within non-transcribed non-coding GSR categories; there are 159 *S. cerevisiae* sequences associated with 148 genes and 99 *S. paradoxus* sequences associated with 97 genes. The most represented GSR categories among these sequences are CDS with 7 associated sequences (all in *S. cerevisiae*) and the UTRs with 18 associated sequences. 39

genes are associated with multiple GSR categories with positive D values (see Table 2.2), with additional support for the 15 genes which are significant at the same GSR category in both species (all non-coding non-transcribed), indicating the persistence of multiple alleles over longer time intervals. Since sequences under balancing selection are sometimes associated with recombination breakpoints (*32*), it is plausible that overrepresentation of the distal non-coding GSR categories in this context may be due in part to the presence of breakpoints rather than selection. Since the historical rate of outcrossing in *S. cerevisiae* is low, estimated at 1 in 50,000 generations (*26*), such signatures should persist within clonal linages.

2.3.4 Polymorphism and divergence of GSR categories

The inclusion of two species in this study allows us to evaluate whether the neutral theory, which predicts a positive correlation between levels of intraspecies polymorphism and interspecies divergence (*27*), applies to patterns of yeast divergence. While Tajima's D test was applied to each GSR for every gene separately, we applied the HKA test (*24*) to each GSR category genome-wide, using 3 variance terms (*S. cerevisiae* polymorphism, *S. paradoxus* polymorphism, and interspecies divergence) computed from all sequences associated with each category. We performed 10 hypothesis tests and rejected the neutral theory for all GSR categories except the 5′ composite ($P < 0.01/10$, Table 2.3) (5′ proximal shows a borderline significance). All rejections are due to large deviations from expectation, except for CDS which shows significantly small deviation ($\chi^2/df = 0.789$). This suggests that at least for CDS most variation, within or between-species, may be deleterious, consistent with a broad role for stabilizing selection. Also the large discrepancy in CDS polymorphic variation between species is consistent with population substructure in *S. cerevisiae*.

Of the 3 variance terms, divergence is the largest term for 8 categories, while *S. cere-*

visiae polymorphism is the largest for 2 categories: 5′ distal and 3′ distal. Also, levels of *S. cerevisiae* polymorphism are greater than levels of *S. paradoxus* polymorphism for all categories except 5′ UTR (though they are nearly identical, 30,564 *vs.* 30,761, respectively). Comparisons of variance terms are not clearly interpretable however, since the number of taxa differ between species. Instead the 3 error terms constituting the HKA statistic indicate that *S. cerevisiae* polymorphism shows the largest deviation from neutral expectation for all GSR categories except intron and 3′ distal, suggesting that excess intraspecies polymorphism contributes the most to rejection of neutrality, rather than interspecies divergence. This is consistent with the overall significant excess of segregating sites in *S. cerevisiae* relative to *S. paradoxus* (Table 2.1).

The intron category shows excess error in *S. paradoxus* polymorphism, consistent with elevated pairwise variation and elevated segregating sites relative to *S. cerevisiae* (Figure 2.1). The intron also exhibits the least error in divergence compared to polymorphism, which is unique among the 10 categories. This apparent reduction in divergence might indicate cross-species conservation of functional intronic elements, including splice junction motifs and other regulatory elements that facilitate spliceosome complex formation (*28, 29*). Nevertheless the intron has the largest χ^2 value relative to degrees of freedom ($\chi^2/df = 1.525$), perhaps suggesting a functional role for the excess *S. paradoxus* polymorphism. In contrast, the 3′ distal category exhibits the greatest error in divergence and the least error in *S. paradoxus* polymorphism, although all 3 error terms are relatively similar. As the only category showing excess divergence this may be surprising, since one could argue that other categories more closely associated with functional elements should be targets of species-specific selection. However, our data do not support this for the CDS category, arguably the category most closely associated with functional elements. Secondly, 5′ and 3′ proximal categories, which are also non-transcribed, exhibit relatively large excesses in

divergence (although not the largest of the 3 error terms). Alternatively, the 3′ distal pattern could be explained as a lack of excess polymorphism in *S. cerevisiae*, or a combination of decreased polymorphism and increased divergence. Overall, we find little evidence for deviation from neutrality due to excess divergence (i.e. species-specific adaptation), however the predominant excess of *S. cerevisiae* polymorphism, in particular for the CDS category, suggests a possible confounding of *S. cerevisiae* demographic effects.

Evolution of GSR categories grouped by GO term

Although we can reject neutrality for 9 out of 10 GSR categories using a genome-wide HKA test, it is possible that levels of polymorphism and divergence for a subset of sequences within a category are consistent with neutrality. To test the neutrality of yeast evolution at a finer resolution, we applied the HKA test to each GSR category grouped by 88 yeast GO Slim terms, which serve as qualitative phenotypic annotations (by proxy of each GSR category's adjacent gene). This resulted in a 10×88 table totaling 880 hypotheses, each of which tests whether a phenotypically-related group of sequences sharing the same structural category appears to evolve neutrally. Of the 851 hypotheses tested (29 hypotheses could not be tested due to small sample size), only 27% were rejected at a 1% false-discovery rate. Thus when grouping sequences by GSR category and by phenotypic annotation, we cannot reject neutrality for the majority of hypotheses. Table 2.4 summarizes this result as the number of GO terms rejected by the HKA test for each GSR category. Averaging over categories, 21.9 out of 76 significant GO terms are rejected, but the number of rejected terms changes substantially across categories. CDS and UTR associate with the most terms (58 and 47 terms) and 5′ proximal, 5′ distal, and intron associate with the fewest (3, 6, and 6 terms). Since GO term annotations are used to group sequences according to some biological phenotype, variation in which may be recognized by natural selection,

we can use the number of terms rejected for a given GSR category as a rough proxy for the relative strength of selection in one category over another. The force of selection on CDS appears to be 8-fold stronger than the force on intron and 20-fold more than on the $5'$ proximal category (*i.e.* promoters).

The fact that the HKA test rejects the neutrality of CDS and UTR categories for the majority of GO terms emphasizes the impact of these categories on phenotypic expression and implies that their alteration can carry a noticeable fitness cost. In contrast, Tajima's D test indicated that proximal categories undergo the strongest purifying selection in both species (Table 2.6), but proximal categories, especially $5'$ proximal, nevertheless lack phenotypic associations. This discrepancy may have arisen because Tajima's D tests whether allele frequencies at a single locus are consistent with neutral mutations (*18*), whereas the HKA test determines whether patterns of polymorphism and divergence correlate across multiple loci (*24*). This suggests that while $5'$ proximal and $5'$ distal GSR categories show patterns of variation similar to the intron when grouped by phenotype, this variation likely sees negative selection. This implies the presence of conserved, though not phenotypically associated, functional elements in these regions (*e.g.* transcription factor binding elements (*30*)).

The common occurrence of relatively elevated non-coding variation in non-transcribed GSR categories (compared to UTRs which are transcribed non-coding) also suggests an alternative possibility of transient mutational hotspots, which could result from recombination breakpoints or locally elevated mutation rates (*31*). Several non-transcribed sequences exhibit positive values of Tajima's D, suggestive of recombination breakpoints in particular (see above). Recombination hotspots are associated with regions of balancing selection (*32*), and our results show that balancing selection is associated with composite, distal, and proximal categories (Table 2.2).

Although recombination hotspots may contribute to the pattern of polymorphic variation, non-transcribed sequences, especially proximal sequences, exhibit strongly negative D values, and the weak pattern of excess divergence in 5′ and 3′ proximal categories (Table 2.1) supports a role for natural selection operating on functional elements in proximal sequences, rather than mutational hotspots. Having subdivided 5′ and 3′ proximal categories by GO term, we also find that 5′ proximal and distal categories each exhibit 5-fold increases in median divergence compared to the median divergence across all GSR categories, and significantly more divergence compared to polymorphism in *S. paradoxus* (t-test, $P = 0.05$). Subsetting GSR categories by phenotypic association thus enhances the pattern of elevated divergence compared to polymorphism seen in the genome-wide HKA tests. (Exceptions include the 5′ proximal GSR category associated with the fundamental cellular processes of protein folding and meiosis and the high-level term cellular component, which are rejected by HKA tests due to an apparent reduction in divergence.) Excess divergence suggests that evolution of transcriptional regulatory elements in the 5′ proximal and distal GSR categories has contributed the most to species divergence (*33, 13*) and could also potentially contribute to the segregation of domestic strains of *S. cerevisiae*. Turnover of functional regulatory elements may even explain the pattern of variation for proximal sequences that do not exhibit elevated divergence, since changes in cis-regulatory elements contributing to gene expression does not necessarily alter the pattern of expression (*34*).

Association of GO terms with non-neutral GSR categories

In the previous section we used the number of GO terms rejected associated with a GSR category as a rough proxy for that category's relative strength of selection. An alternative way to correlate phenotype with GSR category evolution is to associate the number of unique GSR categories rejected by the HKA test with a particular GO term. GO terms

with a large number of non-neutral GSR categories can be seen as requiring multifaceted selection in related gene regions (e.g. on the promoter, 5′ and 3′ UTRs, and CDS), whereas terms with a small number may only require selection on 1 or 2 GSR categories. Thus this number indicates the overall degree of selection associated with a particular term, in both coding and regulatory sequence.

Each GO term is associated with a median of 3 non-neutral GSR categories. The high-level terms cytoplasm and cellular component are the only ones associated with the maximum number of 8 GSR categories, while 36 terms (47%) have only 1 or 2 GSR categories (Table 2.5). This indicates that most biological processes in yeast undergo selection in a small number of structural regions rather than across an entire gene locus. Different metabolic terms associate with either few, some, or many non-neutral GSR categories, while transcriptional and developmental terms (*e.g.* budding, cell cycle, pseudohyphal growth, meiosis, and most generally transcription) associate with numbers of GSR categories at or above the median. Interestingly the 2 terms with the most non-neutral GSR categories (cytoplasm, cellular component) have inherently spatial functions and likely require related mRNAs and proteins to be transported to specific locations within a cell at certain concentrations. Selection operating on both promoters and UTRs is not surprising for such terms that require both transcriptional and post-transcriptional control of gene expression (*35, 36*).

As previously mentioned few GO terms associate with non-neutral promoter and intron sequences (Table 2.4). The 6 GO terms that do associate with non-neutral introns (cytoplasm, DNA binding, DNA metabolic process, membrane, organelle organization and biogenesis, transporter activity) have more than the median number of non-neutral GSR categories (Fisher's Exact test, $P < 0.01$). This is not true for terms associating with non-neutral promoters ($P = 0.13$). This suggests that sequences associated with non-neutral

introns tend to experience selection on multiple genic structural regions, while sequences associated with non-neutral promoters can tolerate neutral evolution on other genic structural regions.

Although the majority of GO terms are associated with GSR categories exhibiting a pattern of variation consistent with purifying selection, 7 terms associate with GSR categories with excess divergence, suggestive of positive selection; 5 terms associate with the 3′ distal category and 2 terms with the 3′ UTR (Table 2.5). Interestingly these latter 2 terms (ribosome and structural molecule activity) share 4 non-neutral GSR categories and show similar ordering of polymorphism and divergence components across all 10 GSR categories ($R = 0.86$), indicating loci labeled by these terms may experience similar selection pressures. This similarity may be explained by the large proportion of genes common to both groups (66.2%, Fisher's Exact test, $P < 10^{-10}$).

Several lines of evidence suggest that one mechanism through which *S. cerevisiae* and *S. paradoxus* have adapted differently is by altered regulation of the 3′ UTRs of ribosomal genes. The ribosomal gene group consists of more than 300 genes which are coordinately expressed, conserved among Ascomycetes, and required for protein synthesis and organism growth during the G_1-phase of the cell cycle (*37, 34*). HKA tests also indicate that the 5′ proximal and distal categories (promoters) of ribosomal genes show reduced divergence consistent with purifying selection (FDR < 0.05 for 5′ distal, suggestive $P < 0.025$ for 5′ proximal), suggesting that both species employ similar transcriptional regulatory dynamics for ribosomal genes. In contrast, evidence for positive selection in the 3′ UTRs of ribosomal genes (see above) suggests that these species differ with respect to post-transcriptional regulation. Since these genes are expressed collectively, altering 3′ UTR sequences may allow the abundance and localization of particular ribosomal genes to be controlled with more precision (*38*). In fact regulation of growth rate through controlled ribosome synthe-

sis is thought to be a mechanism by which budding yeast populations maintain cell size homeostasis (*39*). It is interesting that the complementary set of ribosome biogenesis genes is not associated with excess divergence in any GSR category; rather 5 GSR categories show evidence of purifying selection, including both UTRs.

In summary, association of GO terms with the number of unique, non-neutrally evolving gene regions indicates that most terms only show evidence of selection in 3 or fewer GSR categories. A few terms do show evidence of positive selection in the 3′ distal and 3′ UTR categories, notably implicating altered post-transcriptional regulation of ribosomal genes as one specific mechanism which has diverged between *S. cerevisiae* and *S. paradoxus*.

2.4 Discussion

The recent release of genome-wide polymorphism data for *Homo sapiens* (*40*) has marked the incredible progress in providing affordable genome sequencing for model organisms, while subsequent phenotype association studies pinpointing loci affected by natural selection illustrate the potential of whole-genome population data (*41, 42, 43*). While these analyses typically suffer from low power and poorly understood genome and population structure, similar multi-genome sequencing efforts have also been applied to a much more tractable model organism, the budding yeast, for which it is possible to study genic and population structure in a systematic manner (*12*).

Analysis of polymorphism data for the two yeast species *S. cerevisiae* and *S. paradoxus* has revealed a genome-wide pattern of variation consistent with moderate purifying selection within species and limited divergence between species. *S. cerevisiae* exhibits an excess of low-frequency variation within CDSs and non-transcribed genic structural re-

gions, a pattern not seen in *S. paradoxus*. Thus Tajima's D values are generally more extreme in *S. cerevisiae*, although no genome-wide GSR category deviates significantly from neutrality due to low power. Both Tajima's D and HKA test results report selection to be strongest within CDS, 5′ proximal, and 3′ proximal GSR categories, in a manner consistent with purifying selection. Conservation of yeast genic structural regions appears very similar to that of rodent species, where the coding sequence comes under the strongest purifying selection, followed by relaxed selection on adjacent sequences (*e.g.* the UTRs), and subsequent peaks of selection flanking these (e.g. proximal promoters) on both 5′ and 3′ sides (*44*). Although interspecies divergence appears limited overall, genome-wide and GO term-specific HKA tests reveal a pattern of excess divergence for 5′ and 3′ proximal GSR categories suggesting potentially adaptive species-specific changes in the regulation of gene expression. Moreover, evidence for positive selection in the 3′ UTRs of ribosomal genes suggests species-specific alteration of post-transcriptional regulation associated with control of cellular growth rates. Regulation of growth rate can be used to modulate timing of the developmental transition from vegetative growth to reproduction, a classic life-history strategy (*45*). Different strategies taken by *S. cerevisiae* and *S. paradoxus* may help explain their species-specific mating dynamics (*11*).

This analysis reveals that patterns of intraspecies polymorphism and interspecies divergence may be modulated by the structural/functional context of a given sequence. In particular, we find a clear distinction in the evolution of UTR (transcribed) and proximal (non-transcribed) non-coding GSRs. UTR and proximal categories show comparable average $\hat{\theta}$ (0.041 *vs.* 0.046, t-test, $P > 0.05$) and $\hat{\pi}$ (0.023 *vs.* 0.022, t-test, $P > 0.05$), indicating a similar pattern of genome-wide variation between species. Yet in both species the UTR exhibits the least negative Tajima's D values of any category (-1.26), while the proximal GSR exhibits the most negative values (-1.92) (t-test, $P < 0.05$). Also, there is

no correlation between $\hat{\theta}$ and D across GSR categories ($R = -0.04,\ P = 0.85$), but when considering UTR and proximal categories separately, we find that $\hat{\theta}$ does correlate with D positively in the former situation ($R = 0.75$), but negatively in the latter ($R = -0.96$). Since $\hat{\theta}$ is sensitive to low-frequency variation, differences in the relationship between D and $\hat{\theta}$ between UTR and proximal GSR categories suggest distinct evolutionary dynamics of low-frequency variation. Two explanations appear plausible. Either novel mutations falling within UTRs exist for such a short time they fail to leave a population genetic signature, or mutations in proximal GSRs more effectively create phenotypic variation, which is then eliminated by purifying selection. GO term association identified the UTR as more functionally relevant than the proximal GSR category, supporting the first hypothesis, whereas proximal GSRs exhibit more divergence between species, despite strong purifying selection within species, suggesting rapid turnover of functional cis-elements which might underly gene expression divergence (*46*). Either way, selection on similar patterns of genetic variation differs between UTR and proximal GSR categories, suggesting inherent structural/functional differences.

2.4.1 Confounding selection with demographics

Unlike *S. paradoxus*, whose isolates are predominantly associated with a woodland ecotype, isolates of *S. cerevisiae* come from a variety of environments. In the dataset we analyzed, only 2 were isolated from oak trees, 8 came from vineyards, 7 came from human hosts, and the remaining were either cultured from palm trees, fruits, or soil or are associated with bread making or saké fermentation. The most significant demographic subdivision separates vineyard and saké lineages from the remaining populations (*4, 8*) and likely confounds identification of loci under selection. Thus the combination of excess low-frequency variation within species with excess segregating sites relative to *S. paradoxus*

(particularly among coding sequences) is consistent with the idea of segregating clonal populations undergoing population expansions. Although subsequent partitioning of this polymorphism by ecotype may clarify this picture, additional genomes may be needed to retain the power of large sample sizes. Nevertheless a shared genome architecture, evidence of recent postzygotic reproductive isolation between woodland populations of both species (*5*), and a large population size all argue that purifying selection likely pervades the evolution of *S. cerevisiae* populations as well.

2.4.2 Recombination and linkage disequilibrium

When evaluating the role of natural selection from patterns of genetic variation, the effects of recombination and linkage disequilibrium (LD) must also be considered. Reduced homologous recombination violates an assumption of the evolutionary models employed and could lead to high levels of hitchhiking or background selection, while persistent LD can progressively reduce observed heterozygosity (*47*). The following lines of evidence suggest that these phenomena does not strongly bias our results. First, genomes should recombine if sexual reproduction is typical. Consistent with this, natural budding yeast populations are typically found in the diploid homothallic state in nature, indicating a general ability to reproduce sexually. *S. paradoxus* in particular has consistently shown evidence of a flat, freely recombining sexual population structure, in which variation accumulates at larger spatial scales until constrained by geographic barriers (*7, 6*). This suggests both that recombination is typical in *S. paradoxus* and that reproduction often occurs by outcrossing, resulting in genetic admixture that will prevent loss of heterozygosity. Finally, the HKA test is conservative when the assumption of free interlocus recombination is violated (*24*).

While evidence for recombination and outcrossing within *S. cerevisiae* is less conclusive, woodland and vineyard populations are reported to exhibit relatively short linkage

blocks suggestive of elevated recombination (*8, 22*). While we did not test for LD directly, its presence is suggested by reduced pairwise variation in all regions of *S. cerevisiae* relative to *S. paradoxus* in our data set and prior evidence of a predominantly clonal reproduction (which can generate LD (*31*)) among closely related woodland isolates (*5*). Two pieces of evidence suggest that LD will not strongly affect our results. A rather low rate of outcrossing (1 outcrossing event per 50,000 generations) has been estimated (*26*) from a population with a high inbreeding coefficient, suggesting that variation that may be affected by LD (via clonal reproduction) is nevertheless highly persistent in yeast populations. Secondly, looking at our results, few of the 39 loci associated with balanced polymorphisms over multiple GSR categories lie adjacent or nearly adjacent to each other. Instead these loci are distributed across 14 chromosomes, indicating they are generally unlinked. One exception is a stretch of sequence on chromosome 12 which contains 4 ORFs (YLR154C-H, YLR156C-A, YLR157C-C, YLR159C-A), each of which show evidence of balancing selection across both 3′ proximal and distal GSR categories (Table 2.2). These ORFs are small and orient in the same direction, but they interleave and overlap other ORFs which are not associated with balancing selection. This interleaved pattern of balanced polymorphisms may be explained by persistent heterozygosity maintained in the sampled populations.

2.4.3 Population parameter estimates using neutral loci

Identification of neutrally evolving DNA sequences is necessary to evaluate accurately rates of sequence evolution and the relationships among species. It is generally thought that spliceosomal introns harbor neutral DNA (*17*), and so intronic loci are often used to estimate the rate of accumulation of mutations under neutral drift. The analysis presented here indicates that at least for yeast species, intronic sequences contain predominantly functionally neutral DNA (weak association with GO terms) and a limited number of low-frequency

mutations (neutral Tajima's D within both species). The relative excess polymorphism seen in *S. paradoxus* introns is somewhat surprising, although it is only slightly greater than that seen in *S. cerevisiae*. Also, the intron appears to be the GSR category most compatible with neutrality, as reflected by the least negative Tajima's D value in *S. paradoxus*.

Consequently we used a subset of neutrally evolving intronic sequence data to estimate 252.8×10^3 years as the time since speciation of *S. cerevisiae* and *S. paradoxus*, with corresponding estimates of coalescence within each species as 12.4×10^3 years and 37.4×10^3 years, respectively. Despite estimation of divergence using introns instead of coding sequences and a slightly lower rate of reproduction (6.3 vs. 8 generations per day, see Materials and Methods), our coalescence estimate for the *S. cerevisiae* species is 500 years longer than the the coalescence between vineyard and saké populations reported in (*4*). This strengthens support of all estimates. The divergence time of the entire *Saccharomyces sensu stricto* complex is estimated to be 5–20 million years (*48*). The fact that our estimate of speciation time between *S. cerevisiae* and *S. paradoxus* is at least 20-fold shorter than this suggests that *S. cerevisiae* and *S. paradoxus* might share a closer ecological and evolutionary relationship than has been appreciated.

Effective population size was also estimated using $N = \hat{\pi}/4\mu$. This yielded populations of 16.3×10^6 and 41.7×10^6 organisms for each species, respectively, which are comparable to other unicellular eukaryotes (*49*). This recapitulates the sensitivity of yeast populations to mutations having only minimal fitness cost ($s \approx 3 \times 10^{-8}$), which is likely one reason for the large proportion of the genome under stabilizing selection.

2.5 Materials and Methods

Yeast genome sequences and annotations

The most recent version of the S288c genome (all genomic ORFs), along with annotations delimiting exon boundaries for every open reading frame (ORF), were downloaded from SGD (http://yeastgenome.org). The latest versions of 39 *S. cerevisiae* and 28 *S. paradoxus* genomes were downloaded from NCBI (ftp.ncbi.nih.gov/genomes/), Sanger Institute (www.sanger.ac.uk/Teams/Team118/sgrp/), Broad Institute (http://www.broad.mit.edu/annotation/genome/saccharomyces_cerevisiae/Home.html), and Stanford University Genome Technology Center (http://med.stanford.edu/sgtc/). 5′ and 3′ UTR boundaries were obtained from (*50*).

Annotated ORFs and flanking sequence were parsed into 8 primary and 2 composite GSR categories: Coding sequence (CDS), intron, 5′ UTR, 5′ proximal, 5′ distal, 5′ composite, 3′ UTR, 3′ proximal, 3′ distal, and 3′ composite. Proximal GSRs are defined from the terminal end of a UTR (if known) to 500 nucleotides away from the ORF; if a UTR boundary is not known, the proximal GSR begins after the respective start or stop codon. Distal GSRs are defined as 500 nucleotides from one ORF to the boundary of the next adjacent ORF. Each composite GSR encompasses all contiguous non-coding sequence between two ORFs. For example the 5′ composite GSR for a particular gene is the concatenation of the gene's 5′ distal, 5′ proximal, and 5′ UTR GSRs.

Genome alignments

We wrote a gene identification algorithm to find ORFs within each unannotated genome. This algorithm takes as input a reference genome with corresponding exon, intron, and UTR annotations and an unannotated (target) genome. Following a local alignment of

each reference ORF to a target genome (BLAST-All), candidate ORF boundaries were identified by maximal length concatenation of local, overlapping high-scoring sequence pairs (BLASTN, E-value 1), followed by Needleman–Wunsch global alignment of each reference ORF to the recovered maximal length candidate sequence. CDS sequences were obtained by global alignment of a reference intron to the homologous candidate ORF in a target genome and removal of the aligned target subsequence. A protein-coding locus was retained only if its CDS exhibited proper start and stop codons and if its sequence length was a multiple of 3.

To identify these target ORFs from each unannotated genome, we used all verified, putative, and transposable element ORFs corresponding to two reference genomes: S288c for *cerevisiae* and NRRL Y-17217 *paradoxus*. Since there are no intron or UTR annotations available for *S. paradoxus*, all intron and UTR annotations correspond to the S288c genome. Due to sequence divergence, these intron annotations do not always match actual exon–intron boundaries within homologous *S. paradoxus* ORFs. To identify introns from these ORFs, we modified the reference *S. paradoxus* genome by replacing its intron containing ORFs with homologous ORFs from the reference *S. cerevisiae* genome. We thus used intronic sequences from *S. cerevisiae* when parsing introns from target *S. paradoxus* ORFs. To ensure consistent comparative analysis among genomes, all DNA sequences (including those from the two reference genomes) were obtained as output from this gene parsing algorithm.

Sequence analysis

Two measures of site heterozygosity $\hat{\theta}$ were used to assess sequence variation within species. Estimates of average pairwise sequence variation ($\hat{\pi}$) were obtained for each gene by global alignment of all pairs of a particular region among genomes, taking the average number

of nucleotide differences per base. Segregating sites were obtained by progressive multiple alignment (ClustalW), and subsequent polymorphism estimates were computed using $\hat{\theta} = S/L/\sum_1^{n-1} 1/i$ (*16*). Fixed differences ρ between species were estimated as number of segregating sites reported from global alignment of each gene region after randomly selecting one sequence from *S. cerevisiae* and another from *S. paradoxus*, as described in (*24*). Since the 5′ and 3′ composite regions do not include novel sequence, average statistics reported exclude these regions. Only verified and putative ORFs were used for calculations of polymorphism within species and divergence between species (i.e., transposable elements were excluded).

One important caveat to estimating heterozygosity is that spurious indels can bias $\hat{\pi}$ and $\hat{\theta}$. Because genome assembly errors cannot be distinguished from true insertion-deletion events, we correct for this potential error by treating alignment indels of any length as single mutational events. For example, a 10-nucleotide run of gaps in a multiple alignment is considered a single segregating site. In addition, the proximal/distal boundary of $\pm$500 nucleotides is a rather arbitrary delimiter of a gene's closest non-coding sequence (but see (*30*)). Aligned indels that are adjacent to these boundaries (terminal 5′ end of 5′ proximal and 3′ distal sequences and 3′ end of 5′ distal and 3′ proximal sequences) should be considered spurious, and as such are excluded from estimation of $\hat{\pi}$ and $\hat{\theta}$.

Tests of evolution

Tajima's D statistic was computed by comparing $\hat{\pi}$ and $\hat{\theta}$, as described in (*18*). Significance of Tajima's D values was assessed genome-wide using FDR < 0.05 separately for all 10 GSR categories within both species. The HKA statistic was computed by comparing estimates of $\hat{\theta}$ for each species to ρ, as described in (*24*). Significance of the HKA statistic was assessed using FDR < 0.05 separately for each GSR category, stratified by 88 yeast

GO Slim terms (880 total hypotheses). ORFs may be annotated with multiple GO terms and so may be included in more than one hypothesis test.

Population parameter estimates

The divergence time t (in generations) was obtained within and between *S. cerevisiae* and *S. paradoxus* using $k = 2\mu t$, which assumes a constant nucleotide substitution rate ρ since speciation, and a fixed rate of mutation per nucleotide per generation μ. We used 1.84×10^{-10} as the mutation rate (*4*). We used $\hat{\pi}$ and ρ as the substitution rates to estimate within species coalescence times and speciation time, respectively. For the former estimate, we used a subset of intronic loci whose Tajima's D values were bounded by $(-1,1)$ to ensure neutrality. A total of 36, 57, and 164 loci were used to estimate coalescence times within *S. cerevisiae*, within *S. paradoxus*, and between species, respectively. All intronic loci were used for the interspecies estimate since Tajima's D only applies to polymorphic data. To convert t into years, we estimated the number of generations per day based cycle length estimates obtained from α-factor synchronization of 10 haploid woodland isolates of both yeast species at an environmentally realistic temperature of 18°C. We estimated that *Saccharomyces* reproduces every 227.92 min, or 6.32 generations per day. This estimate is slightly lower than 8 generations per day, which was used by (*4*). Effective population size N was estimated from the same data using $N = \hat{\pi}/4\mu$, where $\hat{\pi}$ and μ are described above.

Data Availability

All relevant DNA sequence data, multiple alignments, and statistics can be accessed through the Budding Yeast Gene Evolution database (http://yeastpopgenomics.org).

Acknowledgments

We wish to thank Dr. Paul Sniegowski and Dr. Warren Ewens for thorough readings of this manuscript as well as members of the Kim lab for helpful comments and advice. We would also like to acknowledge the Sanger Institute and the laboratory of Dr. Ed Louis for making their yeast genome sequence collection publicly available at an early stage.

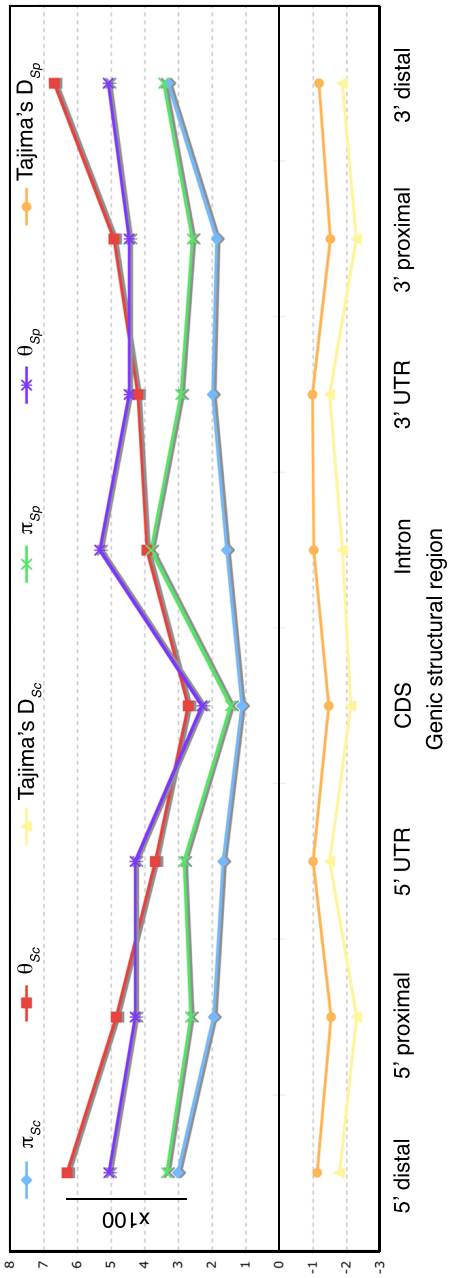


Figure 2.1: Genome-wide distribution of $\hat{\pi}$, $\hat{\theta}$, and Tajima's D across GSR categories $\hat{\pi}$ and $\hat{\theta}$ estimates have been multiplied by 100. Numerical values are presented in Table 2.6.

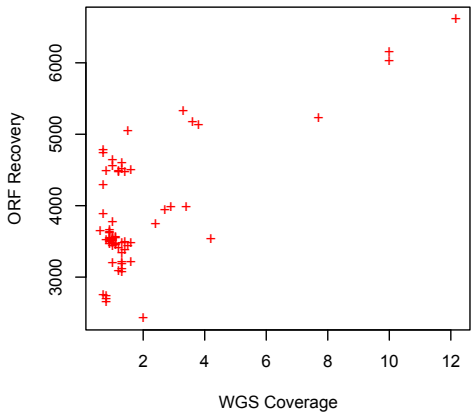


Figure 2.2: Scatterplot of number of ORFs recovered by the genome identification algorithm versus the whole-genome shotgun coverage of each genome

To determine which factors affect gene recovery from an assembled genome sequence, gene count (number of recovered ORFs) was regressed with WGS coverage and species identity. Both features are significant (each $P < 0.001$) and together explain 70% of the variation, but 49.6% of the variation is explained by WGS coverage. Thus quality of experimental data is the primary determinant to this algorithm's performance.

Table 2.1: Differences in estimates of site heterozygosity and Tajima's D between species

Category	$t_{\hat{\theta}}$	(p-value)	$t_{\hat{\pi}}$	(p-value)	t_D	(p-value)
5′ composite	10.1750*	($< 10^{-5}$)	-18.7029*	($< 10^{-5}$)	-7.2788*	($< 10^{-5}$)
5′ distal	11.2143*	($< 10^{-5}$)	-2.9574	(0.0016)	-5.4714*	($< 10^{-5}$)
5′ proximal	11.0637*	($< 10^{-5}$)	-19.3665*	($< 10^{-5}$)	-49.0574*	($< 10^{-5}$)
5′ UTR	-5.9552*	($< 10^{-5}$)	-23.6910*	($< 10^{-5}$)	-17.7960*	($< 10^{-5}$)
CDS	16.1338*	($< 10^{-5}$)	-25.5250*	($< 10^{-5}$)	-14.9104*	($< 10^{-5}$)
Intron	-3.7128*	(0.0001)	-8.9084*	($< 10^{-5}$)	-10.6384*	($< 10^{-5}$)
3′ UTR	-2.0783	(0.0189)	-13.3272*	($< 10^{-5}$)	-15.3247*	($< 10^{-5}$)
3′ proximal	7.0091*	($< 10^{-5}$)	-11.0176*	($< 10^{-5}$)	-19.0603*	($< 10^{-5}$)
3′ distal	10.9762*	($< 10^{-5}$)	-0.8140	(0.2080)	-10.1872*	($< 10^{-5}$)
3′ composite	8.2912*	($< 10^{-5}$)	-15.6119*	($< 10^{-5}$)	-14.0071*	($< 10^{-5}$)

t-tests are performed between species for each GSR category ($Sc - Sp$). *indicates significance using $P < 0.01/10$.

Table 2.2: Genes with multiple GSRs under balancing selection

	GSR categories	
Gene	*S. cerevisiae*	*S. paradoxus*
YAR019C	3′ distal	3′ distal
YAR068W	3′ proximal, 3′ distal	
YBL071W-A	CDS, 3′ distal	
YBR301W	5′ composite	5′ composite, 5′ proximal, 3′ proximal
YDL047W	3′ distal	3′ distal
YDL240W	3′ distal	3′ distal
YDL244W	5′ composite, 5′ distal	
YDR542W	5′ composite, 5′ proximal	
YER189W	5′ composite, 5′ distal	
YFL064C	5′ composite, 5′ distal	
YGR295C	5′ composite, 5′ proximal, 3′ proximal	
YHL046C	5′ composite, 5′ proximal	
YHL049C	5′ composite, 5′ proximal	
YHR008C	5′ distal	5′ distal
YIL172C	5′ composite, 5′ proximal	
YIR039C	5′ composite, 5′ proximal	
YJL163C	3′ distal	5′ distal, 3′ distal
YJL221C	5′ composite, 5′ proximal	
YJL223C	5′ composite, 5′ distal	
YJR156C	5′ distal, 5′ composite	
YKL140W	3′ distal, 5′ distal	3′ distal
YKL141W	3′ distal	3′ distal
YKL151C	3′ distal	3′ distal
YLL041C	3′ distal	3′ distal
YLR004C	5′ distal	5′ distal
YLR154C-H	3′ distal, 3′ proximal	
YLR156C-A	3′ distal, 3′ proximal	
YLR157C-C	3′ distal, 3′ proximal	
YLR159C-A	3′ distal, 3′ proximal	
YLR461W	5′ composite, 5′ proximal	
YNL033W	3′ proximal, 3′ distal	
YOL157C	5′ composite, 5′ proximal, 3′ proximal	
YOL161C	5′ composite, 5′ proximal	
YOR388C	5′ distal	5′ distal
YPL017C	5′ distal	5′ distal
YPL050C	5′ distal	3′ proximal
YPR041W	3′ distal	3′ distal
YPR080W	5′ distal, 3′ proximal	
YPR202W	5′ composite, 5′ proximal	

Table 2.3: HKA statistics for each GSR category

Rank	Category	χ^2	(p-value)	df†	S_{Sc}	S_{Sp}	D	χ^2_{Sc}	χ^2_{Sp}	χ^2_D
9	5′ composite	9,531.15	(0.0611)	9,746	635,743	474,875	**653,839**	**3,795.70**	2,670.27	3,065.18
6	5′ distal	5,168.52*	($< 10^{-5}$)	4,518	**272,114**	172,920	252,856	**1,851.37**	1,650.95	1,666.20
8	5′ proximal	9,968.36*	($< 10^{-5}$)	9,530	308,785	253,504	**348,201**	**3,945.21**	2,672.98	3,350.17
2	5′ UTR	6,197.23*	($< 10^{-5}$)	4,322	30,564	30,761	**46,406**	**2,456.11**	1,765.70	1,975.42
10	CDS	8,147.17*	($< 10^{-5}$)	10,320	756,991	547,739	**764,388**	**3,340.01**	2,096.81	2,710.35
1	Intron	457.64*	($< 10^{-5}$)	300	8,194	8,154	**9,168**	152.57	**165.96**	139.11
3	3′ UTR	6,051.23*	($< 10^{-5}$)	4,294	37,429	34,108	**48,787**	**2,439.73**	1,635.95	1,975.55
5	3′ proximal	11,458.70*	($< 10^{-5}$)	9,260	233,814	200,544	**284,303**	**4,478.39**	3,243.73	3,736.54
4	3′ distal	3,450.69*	($< 10^{-5}$)	2,676	**165,602**	99,160	145,827	1,200.40	1,027.05	**1,223.24**
7	3′ composite	11,123.00*	($< 10^{-5}$)	9,764	453,878	354,039	**479,475**	**4,311.66**	3,248.24	3,563.11
	Average	5,112.44	($< 10^{-4}$)	5,653	226,687	168,361	**237,492**	**2,482.97**	1,782.39	2,097.07

Averages exclude 5′ and 3′ composite statistics. †denotes the degrees of freedom for each application of the HKA test. *indicates significance using FWER $< 0.01/10$. The first column ranks the deviation from neutrality relative to degrees of freedom, χ^2/df, values from greatest to least. Bold values indicate the largest term (of variation and of error) for each GSR category.

Table 2.4: Number of GO terms rejected by the HKA test within each GSR category

Category	GO Terms
5′ composite	6
5′ distal	6
5′ proximal	3
5′ UTR	47
CDS	58
Intron	6
3′ UTR	43
3′ proximal	24
3′ distal	17
3′ composite	9

Table 2.5: Number of GSR categories rejected by the HKA test, grouped by GO term

GO Slim Term	Categories	GO Slim Term	Categories
cytoplasm	8	vesicle-mediated transport	2
cellular component[a]	8	site of polarized growth	2
All	7	signal transducer activity	2
translation	6	response to chemical stimulus	2
structural molecule activity[b]	6	phosphoprotein phosphatase activity	2
ribosome[b]	6	oxidoreductase activity	2
plasma membrane	6	lyase activity	2
organelle organization and biogenesis	6	ligase activity	2
DNA metabolic process	6	endomembrane system	2
transport	5	cytoplasmic membrane-bound vesicle	2
transcription regulator activity	5	conjugation	2
ribosome biogenesis and assembly	5	cofactor metabolic process	2
protein modification process	5	carbohydrate metabolic process	2
nucleus	5	biological process[a]	2
mitochondrion	5	amino acid and derivative metabolic process	2
membrane	5	vitamin metabolic process	1
cell wall organization and biogenesis	5	translation regulator activity	1
transporter activity	4	sporulation	1
transferase activity	4	protein catabolic process	1
transcription	4	peroxisome	1
protein folding	4	peptidase activity[a]	1
molecular function[a]	4	nucleotidyltransferase activity	1
hydrolase activity	4	nuclear organization and biogenesis	1
helicase activity	4	motor activity	1
RNA metabolic process	4	mitochondrial envelope	1
RNA binding	4	microtubule organizing center	1
DNA binding	4	membrane organization and biogenesis	1
signal transduction	3	lipid metabolic process	1
response to stress	3	isomerase activity[a]	1
protein kinase activity	3	heterocycle metabolic process	1
protein binding	3	generation of precursor metabolites and energy	1
other	3	cytoskeleton organization and biogenesis	1
nucleolus	3	cytoskeleton	1
meiosis	3	cytokinesis	1
endoplasmic reticulum	3	cell cortex	1
chromosome	3	cell budding	1
cellular bud	3	Golgi apparatus	1
cell cycle	3		
anatomical structure morphogenesis	3		

Significance of χ^2-tests was evaluated using FDR < 0.05. [a]indicates excess divergence in the 3′ distal GSR category. [b]indicates excess divergence in the 3′ UTR.

Table 2.6: Polymorphism statistics for GSR categories

A) *S. cerevisiae* polymorphism

Category	GSRs	D	(p-value)	Low D	High D	Taxa	$\hat{\theta}$	$\hat{\pi}$	Length
5′ composite	6533.0	-1.9138	(0.0582)	3435.0	18.0	24.5582	0.052	0.0224	705.0561
5′ distal	3687.0	-1.8032	(0.0675)	1908.0	42.0	24.5582	0.0629	0.03	525.3191
5′ proximal	6503.0	-2.3100	(0.0337)	5050.0	13.0	24.5582	0.0483	0.0194	377.231
5′ UTR	2466.0	-1.5091	(0.1341)	54.0	2.0	24.5582	0.0367	0.0166	102.9371
CDS	6575.0	-2.1378	(0.0449)	4680.0	7.0	24.5582	0.0269	0.011	1274.8938
Intron	271.0	-1.8859	(0.0946)	103.0	0.0	30.2694	0.0392	0.0155	241.9446
3′ UTR	2453.0	-1.5068	(0.1291)	83.0	8.0	24.5582	0.0419	0.0197	111.5842
3′ proximal	6365.0	-2.3009	(0.0403)	4606.0	36.0	24.5582	0.049	0.0185	312.4112
3′ distal	2561.0	-1.8842	(0.0595)	1472.0	50.0	24.5582	0.0667	0.0329	514.4155
3′ composite	6534.0	-1.8949	(0.0748)	2520.0	27.0	24.5582	0.0514	0.023	543.2053
Average	3860.125	-1.9172	(0.0755)	2244.5	19.75	25.2721	0.0465	0.0205	432.5921

B) *S. paradoxus* polymorphism

Category	GSRs	D	(p-value)	Low D	High D	Taxa	$\hat{\theta}$	$\hat{\pi}$	Length
5′ composite	4869.0	-1.3008	(0.1881)	16.0	3.0	17.9582	0.0452	0.0284	654.6423
5′ distal	2516.0	-1.1395	(0.1765)	118.0	40.0	17.9582	0.0505	0.0333	485.6484
5′ proximal	4844.0	-1.5522	(0.1494)	536.0	1.0	17.9582	0.0427	0.0262	362.4393
5′ UTR	2163.0	-1.0167	(0.2248)	16.0	2.0	17.9582	0.0427	0.0282	102.2911
CDS	5250.0	-1.4834	(0.1612)	36.0	0.0	17.9582	0.0228	0.0143	1408.0124
Intron	151.0	-1.0366	(0.2855)	2.0	0.0	18.1939	0.0532	0.038	273.7483
3′ UTR	2146.0	-0.9987	(0.2196)	20.0	6.0	17.9582	0.0445	0.0291	111.4171
3′ proximal	4713.0	-1.5307	(0.1466)	721.0	20.0	17.9582	0.0445	0.0257	290.1292
3′ distal	1613.0	-1.1924	(0.1635)	49.0	30.0	17.9582	0.0507	0.0342	484.2899
3′ composite	4877.0	-1.1485	(0.1958)	21.0	12.0	17.9582	0.0455	0.029	488.1825
Average	2924.5	-1.2438	(0.1909)	187.25	12.375	17.9877	0.0439	0.0286	439.747

The number of genes represented in each GSR category was determined by counting homologous sequences identified in at least two genomes. p-values for Tajima's D were computed using a Beta distribution. The number of significantly low and high GSRs was determined using FDR < 0.05. Averages exclude 5′ and 3′ composite GSR category statistics.

Table 2.7: Divergence statistics for GSR categories

Category	GSRs	Taxa	$\hat{\pi}$	$\bar{S}$	Identities	Gaps	Length
5′ composite	5261.0	17.9582	0.2681	124.4901	445.4744	11.8055	569.9645
5′ distal	2316.0	17.9582	0.2294	49.8032	150.0104	4.3178	199.8136
5′ proximal	4834.0	17.9582	0.2049	67.2002	255.625	6.08	322.8252
5′ UTR	2314.0	17.9582	0.2827	8.8446	33.0399	1.0264	41.8845
CDS	5203.0	17.9582	0.1099	146.0566	1182.1797	3.2614	1328.2363
Intron	164.0	18.1939	0.2145	58.2667	203.2485	5.5697	261.5152
3′ UTR	2299.0	17.9582	0.274	9.2975	36.1805	1.2046	45.478
3′ proximal	4702.0	17.9582	0.2148	55.0389	193.4956	5.0435	248.5346
3′ distal	1373.0	17.9582	0.2347	28.4776	83.4614	2.277	111.939
3′ composite	5253.0	17.9582	0.2651	91.1438	319.656	8.9713	410.7998
Average	2900.625	17.9877	0.2206	52.8732	267.1551	3.5975	320.0283

The number of genes represented in each GSR category was determined by counting homologous sequences identified in at least two genomes. Averages exclude 5′ and 3′ composite GSR category statistics. $\bar{S}$ is the average number of segregating sites among GSRs in a particular category.

Table 2.8: Controls for the genome identification algorithm

	Genes				Nucleotides		
Genome	Recovered	Total	Differing	% Error	Nonidentities	Average	Median
S288c	6612	6719	13	0.20	28	2.15	2
RM11-1a	6023	6719	49	0.81	2402	49.02	4
YJM789	6162	6719	82	1.33	7708	94.00	5
Y-17217	5227	8171	157	3.00	16828	107.18	106

To evaluate how well our gene identification algorithm identifies ORFs from an assembled genome, the ORFs returned by the algorithm were compared to known ORFs from the same genome for three *S. cerevisiae* strains (S288c, YJM789, and RM11-1a) and one *S. paradoxus* strain (NRRL Y-17217). The proportion of recovered genes is comparable among the *S. cerevisiae* samples and roughly 30% lower in the *S. paradoxus* sample, whose 1452 additional ORFs are likely to be spurious, short-length ORFs or noncoding ORFs. Such sequences would be eliminated by the algorithm due to insufficient alignment power or failure to conform to rules of a protein-coding sequence. Also only 61% of intron genes were identified in *S. paradoxus*, thus excluding an additional 106 known genes. Since no intron annotations exist for *S. paradoxus*, known intronic *S. cerevisiae* genes were used in place of the corresponding *S. paradoxus* genes (see Materials and Methods). Taking these factors into account, sensitivity on the *S. paradoxus* sample is comparable to the *S. cerevisiae* samples.

The number of genes whose known sequence differs from the recovered sequence was also computed to assess the algorithm's accuracy for properly delimiting gene boundaries. *S. cerevisiae* samples had fewer than 100 differing sequences, while *S. paradoxus* yielded 157 differing sequences. Some of the nucleotide differences between the reference and output sets can be attributed to spurious sequence differences between a reference ORF set

and the corresponding input genome. Notably this is the case for the S288c genome, where all differences are single base mismatches (not shown).

The discrepancy between average and median lengths for RM11-1a and YJM789 illustrates that there are two classes of differing genes: one with simple mismatch errors and another with longer, single gap-run discrepancies (not shown). Comparable average and median lengths for the Y-17217 genome suggest dominance of the latter class. The vast majority of these gap-runs occur in the terminal 5′ end of a gene, which would be expected if there were proximal upstream start codons with sufficient intervening sequence homology. In this case BLAST alignment could return a corresponding HSP subsequence, which would be incorporated into a new gene sequence.

Identification of upstream ORFs (uORF) has been a notable issue in annotating the S288c genome, and comparative genomic evidence suggests that the majority of uORFs within 150 bases upstream of a known start codon are real (*51, 52*). Postulated mechanisms for uORF function include translational control through specific mis-sense mutations or length of an uORF subsequence (*53*). Such sequence-level features should effectively maintain an elevated level of evolutionary constraint. Thus although potential uORFs are treated as coding sequence in this paper, such relatively short, conserved sequences should add negligible bias to the summary statistics presented. For *S. paradoxus* in particular, it is possible some sequences represent unannotated introns. Regardless, the lengths of these differing sequences are consistent with either uORFs or introns.

Table 2.9: *S. cerevisiae* polymorphism statistics using genes common between species

Category	GSRs	D	(p-value)	Low D	High D	Taxa	$\hat{\theta}$	$\hat{\pi}$	Length
5′ composite	5260.0	-1.9996	(0.0616)	2607.0	6.0	23.7998	0.0494	0.0219	666.2458
5′ distal	2647.0	-1.8101	(0.0764)	1232.0	33.0	23.7998	0.0635	0.0303	517.2999
5′ proximal	5232.0	-2.2904	(0.0355)	4020.0	5.0	23.7998	0.0480	0.0192	361.1912
5′ UTR	2312.0	-1.5154	(0.134)	50.0	2.0	23.7998	0.0367	0.0166	102.7197
CDS	5265.0	-2.1960	(0.0348)	4019.0	5.0	23.7998	0.0280	0.0115	1376.9718
Intron	165.0	-1.9492	(0.0839)	67.0	0.0	30.2848	0.0425	0.0168	264.0970
3′ UTR	2299.0	-1.5071	(0.128)	79.0	8.0	23.7998	0.0419	0.0198	111.5080
3′ proximal	5091.0	-2.2879	(0.043)	3630.0	26.0	23.7998	0.0481	0.0175	286.3606
3′ distal	1609.0	-1.7662	(0.0732)	812.0	39.0	23.7998	0.0667	0.0340	503.0973
3′ composite	5251.0	-1.8625	(0.0818)	1785.0	15.0	23.7998	0.0479	0.0221	481.8073
Average	3077.5	-1.9153	(0.0761)	1738.625	14.75	24.6104	0.0469	0.0207	440.4057

It is possible that the 1310 genes not identified in *S. paradoxus* may have species-specific properties. Since *S. paradoxus* is highly enriched for verified *S. cerevisiae* genes, this *S. cerevisiae*-specific set is likely enriched for putative genes. To demonstrate that comparisons of polymorphism between species are not consequently biased, diversity measures were recomputed for *S. cerevisiae* using the 5265 genes common among both species. There is a negligible increase in average variability for both $\hat{\theta}$ (0.0006) and $\hat{\pi}$ (0.0002) using the common gene set. Region deviations are also small, the largest being 0.0153 for the 3′ full $\hat{\theta}$ and 0.0011 for the 3′ outer $\hat{\pi}$. Both measures correlate extremely well across regions (0.98 and 0.95, respectively). Tajima's D values also correlate strongly across regions (0.98). The average deviation is 0.0019 in the direction of the complete gene set. 3′ outer shows the largest (but insignificant) discrepancy in Tajima's D of 0.0324 towards the complete gene set.

The number of gene regions called significant within species correlates strongly (0.99 for Low calls and 0.97 for High calls), however the total number of significant genes is sensitive to gene set used. The reduced gene set yields 31.5% fewer significant high genes and 23.5% fewer low genes. While this discrepancy appears to account for the difference in total significantly high genes between species, *S. cerevisiae* nevertheless exhibits a surplus of total low genes compared to *S. paradoxus* when using the common gene set. In summary inclusion of unique *S. cerevisiae* genes in analysis of polymorphism shows a negligible effect on all statistics reported.

Table 2.10: Polymorphism statistics for transposable elements in *S. cerevisiae*

Category	GSRs	D	(p-value)	Low D	High D	Taxa	$\hat{\theta}$	$\hat{\pi}$	Length
5′ composite	88.0	-1.862	(0.0995)	16.0	0.0	36.6932	0.1734	0.073	573.3523
5′ distal	88.0	-1.795	(0.1149)	16.0	0.0	36.6932	0.1778	0.1002	469.25
5′ proximal	88.0	-2.0131	(0.0868)	24.0	0.0	36.6932	0.0802	0.0449	488.2308
CDS	88.0	-2.4395	(0.0368)	42.0	0.0	36.6932	0.0184	0.0034	1847.6852
3′ proximal	88.0	-2.7493	(0.0363)	52.0	0.0	36.6932	0.0469	0.017	485.0571
3′ distal	88.0	-2.1872	(0.0458)	42.0	1.0	36.6932	0.1025	0.0445	482.1875
3′ composite	88.0	-2.0422	(0.0692)	38.0	1.0	36.6932	0.1033	0.0308	729.9773
Average	88.0	-2.2368	(0.0641)	35.2	0.2	36.6932	0.0852	0.042	754.4821

Transposable elements (TE) constitute a family of five classes of retrotransposons totaling 89 ORFs in the *S. cerevisiae* genome. Expression of these elements is regulated predominantly post-transcriptionally (*54*). To understand whether the pattern of sequence variation and evolutionary constraint is unique among these elements, polymorphism and Tajima's D statistics were computed for 88 TEs. (Only one corresponding TE was identified in the *S. paradoxus* genome, see Results.)

Both measures of diversity are nearly twice as large as those corresponding to endogenous protein-coding genes (0.042 vs. 0.021 for $\hat{\pi}$, 0.354 vs. 0.176 for $\hat{\theta}$). Consequently Tajima's D values are more negative among TEs, averaging -2.24, indicating pervasive influence of purifying selection across each locus. Whereas the 5′ proximal, 3′ proximal, and CDS were most negative among endogenous genes, CDS, 3′ proximal, and 3′ distal are significant among TEs. While insignificance at the promoter region reinforces the lack of TE regulation at the transcriptional level, the extremely negative Tajima's D value of -2.75 corresponding to the 3′ proximal likely reflects its role as the site of post-transcriptional regulation. In addition Tajima's D for the 3′ distal region is very negative, indicating selection might also operate on the 3′ direct repeat terminal ends of TEs, possibly inhibiting DNA reintegration.

References

1. Mortimer, R. K. *Genome Res* **10**, 403–409 (2000).

2. Tawfik, O. W., Papasian, C. J., Dixon, A. Y., and Potter, L. M. *J Clin Microbiol* **27**(7), 1689–1691 July (1989).

3. Naumov, G. I., Naumova, E. S., and Sniegowski, P. D. *Can J Microbiol* **44**, 1045–1050 (1998).

4. Fay, J. C. and Benavides, J. A. *PLoS Genet* **1**(1), 66–71 (2005).

5. Kuehne, H. A. *The Genetic Structure and Biogeography of Natural Saccharomyces Populations*. PhD thesis, University of Pennsylvania, (2005).

6. Kuehne, H. A., Murphy, H. A., Francis, C. A., and Sniegowski, P. D. *Curr Biol* **17**(5), 407–411 (2007).

7. Koufopanou, V., Hughes, J., Bell, G., and Burt, A. *Philos Trans R Soc Lond B Biol Sci* **361**(1475), 1941–1946 (2006).

8. Aa, E., Townsend, J., Adams, R., Nielsena, K., and Taylor, J. *FEMS Yeast Res* **6**, 702–715 (2006).

9. Johnson, L. J., Koufopanou, V., Goddard, M. R., Hetherington, R., Schäfer, S. M., and Burt, A. *Genetics* **166**, 43–52 January (2004).

10. Sniegowski, P. D., Dombrowski, P. G., and Fingerman, E. *FEMS Yeast Res* **1**(4), 299–306 (2002).

11. Murphy, H. A., Kuehne, H. A., Francis, C. A., and Sniegowski, P. D. *Biol Lett* **2**(4), 553–556 (2006).

12. Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N., Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R., and Louis, E. J. *Nature* **458**(7236), 337–341 (2009).

13. Haerty, W. and Singh, R. S. *Mol Biol Evol* **9**, 1707–1714 (2006).

14. Fingerman, E. G., Dombrowski, P. G., Francis, C. A., and Sniegowski, P. D. *Yeast* **20**(9), 761–770 (2003).

15. Watterson, G. A. *Theor Popul Biol* **7**(2), 256–276 (1975).

16. Gillespie, J. H. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, second edition, (2004).

17. Chamary, J.-V. and Hurst, L. D. *Mol Biol Evol* **21**(6), 1014–1023 (2004).

18. Tajima, F. *Genetics* **123**, 585–595 November (1989).

19. Smith, J. M. and Haigh, J. *Genetics Research* **23**, 23–25 (1974).

20. Charlesworth, B., Morgan, M. T., and Charlesworth, D. *Genetics* **134**(4), 1289–1303 (1993).

21. Stajich, J. E. and Hahn, M. W. *Mol Biol Evol* **22**(1), 63–73 (2005).

22. Schacherer, J., Shapiro, J. A., Ruderfer, D. M., and Kruglyak, L. *Nature* **458**(7236), 342–345 (2009).

23. Ohta, T. *Nature* **246**, 96–98 (1973).

24. Hudson, R. R., Kreitman, M., and Aguadé, M. *Genetics* **116**, 153–159 May (1987).

25. Strobeck, C., Smith, J. M., and Charlesworth, B. *Genetics* **82**, 547–558 March (1976).

26. Ruderfer, D. M., Pratt, S. C., Seidel, H. S., and Kruglyak, L. *Nat Genet* **38**(9), 1077–1081 (2006).

27. Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, (1983).

28. Howe, K. J. and Ares, M. J. *Proc Natl Acad Sci U S A* **94**(23), 12467–12472 (1997).

29. Libri, D., Lescure, A., and Rosbash, M. *RNA* **6**(3), 352–368 (2000).

30. Harbison, C. T., Gordon, D. B., Lee, T. I., et al. *Nature* **431**(7004), 99–104 (2004).

31. Fay, J. C. and Benavides, J. A. *Genetics* **170** August (2005).

32. Strobeck, C., Smith, J. M., and Charlesworth, B. *Genetics* **82**(3), 547–558 (1976).

33. Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. *Mol Syst Biol* **4**, 159 (2008).

34. Tanay, A., Regev, A., and Shamir, R. *Proc Natl Acad Sci U S A* **102**(20), 7203–7208 (2005).

35. Ringner, M. and Krogh, M. *PLoS Comput Biol* **1**(7), e72 (2005).

36. Shalgi, R., Lapidot, M., Shamir, R., and Pilpel, Y. *Genome Biol* **6**(10), R86 (2005).

37. Gasch, A. P., Spellman, P. T., Kao, C. M., et al. *Mol Biol Cell* **11**(12), 4241–4257 (2000).

38. Wickens, M., Bernstein, D. S., Kimble, J., and Parker, R. *Trends Genet* **18**(3), 150–157 (2002).

39. Jorgensen, P., Rupes, I., Sharom, J. R., Schneper, L., Broach, J. R., and Tyers, M. *Genes Dev* **18**(20), 2491–2505 (2004).

40. Frazer, K. A., Ballinger, D. G., Cox, D. R., et al. *Nature* **449**(7164), 851–861 (2007).

41. Tishkoff, S. A., Reed, F. A., Ranciaro, A., et al. *Nat Genet* **39**(1), 31–40 (2007).

42. Sulem, P., Gudbjartsson, D. F., Stacey, S. N., et al. *Nat Genet* **39**(12), 1443–1452 (2007).

43. Sabeti, P. C., Schaffner, S. F., Fry, B., et al. *Science* **312**(5780), 1614–1620 (2006).

44. Keightley, P. D., Lercher, M. J., and Eyre-Walker, A. *PLoS Biol* **3**(2) February (2005).

45. Gadgil, M. and Bossert, W. H. *Amer Nat* **104**(935), 1–24 (1970).

46. Wittkopp, P. J., Haerum, B. K., and Clark, A. G. *Nat Genet* **40**(3), 346–350 (2008).

47. Fay, J. C. and Wu, C. I. *Genetics* **155**(3), 1405–1413 (2000).

48. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. *Nature* **423** May (2003).

49. Lynch, M. and Conery, J. S. *Science* **302**(5649), 1401–1404 (2003).

50. David, L., Huber, W., Granovskaia, M., et al. *Proc Natl Acad Sci U S A* **103**(14), 5320–5325 (2006).

51. Cvijovic, M., Dalevi, D., Bilsland, E., Kemp, G. J. L., and Sunnerhagen, P. *BMC Bioinformatics* **8**, 295 (2007).

52. Zhang, Z. and Dietrich, F. S. *Curr Genet* **48**(2), 77–87 (2005).

53. Vilela, C. and McCarthy, J. E. G. *Mol Microbiol* **49**(4), 859–867 August (2003).

54. Farabaugh, P. J. *J Biol Chem* **270**(18), 10361–10364 May (1995).

Chapter 3

Heterochronic Evolution Reveals Modular Timing Changes in Budding Yeast Transcriptomes

3.1 Abstract

Previous studies on the evolution of genome-wide gene expression support the hypothesis that transcriptome evolution is limited by stabilizing selection, suggesting that it is difficult for an organism to acquire adaptive, functional changes in gene regulation. However gene expression is a dynamic trait, whose timing is regulated by a complex, polygenic combination of factors. In addition, evolutionary modifications to the architecture of gene regulation have the potential to dramatically alter a gene's expression timing without greatly affecting its average expression level. To evaluate the extent of potentially functional timing changes, the mode of time-dependent transcriptome evolution, and the architecture of timing control, we investigated the evolution of genome-wide gene expression as a dynamical system using transcriptome measurements throughout the mitotic cell-division cycle of budding yeast, for 8 woodland and 1 laboratory strain of *S. cerevisiae* and 1 outgroup of *S. paradoxus*. Despite evidence of strong stabilizing selection on expression levels, most genes show significant evolutionary divergence in expression dynamics at all scales of transcriptome organization, suggesting broad potential for functional timing changes. A

model involving timing changes explains 61% of the between-genome variation in expression dynamics, suggesting that the major mode of transcriptome evolution entails changes in timing (heterochrony) rather than changes in levels (heterometry) of expression. Analysis of heterochrony patterns suggests that timing control is organized into distinct, coherent, and dynamically-autonomous modules. We propose that widespread divergence in expression dynamics may be explained by pleiotropic changes in modular timing control, perhaps mediated by relatively few transcription factors. Gene regulation may utilize a general architecture comprised of multiple discrete event timelines, whose superposition could produce combinatorial complexity in timing patterns.

3.2 Introduction

Recent evolutionary studies using natural and inbred *Drosophila* and *C. elegans* lines have shown that genome-wide gene expression levels are much more conserved than expected by independent measurements of mutational input (*1, 2, 3*), supporting the hypothesis that transcriptome evolution is dominated by stabilizing selection. This implies that organisms should have difficulty acquiring adaptive, functional changes in gene regulation, mediated either by changes in the ability of transcriptional regulatory factors (TFs) to bind DNA motifs (changes in regulatory interactions) or by changes in the expression levels of TFs themselves. Since gene regulation involves highly connected cascades of TFs (*4, 5, 6*), both kinds of regulatory change may be limited due to the broad potential for negative pleiotropic consequences (*7*). Given this broad potential for deleterious changes in gene regulation, how do organisms achieve adaptive functional divergence of gene expression?

While gene expression is a complex, quantitative trait, it is also a dynamic trait, such that the timing of gene expression is regulated by a complex, polygenic combination of

factors (*4, 8, 9, 10, 11*). One possibility is that functional changes may occur via changes in the timing of gene expression, inducing temporal shifts in the expression trajectories of some genes relative to others (heterochrony) (*12, 13*). Moreover, evolutionary modifications of gene regulation have the potential to dramatically alter a gene's expression timing without greatly affecting its average expression level (*14, 15*). In this study, we investigated the evolution of genome-wide gene expression as a dynamical system, to evaluate the extent of potentially functional timing changes, the mode of time-dependent transcriptome evolution, and the architecture of timing control. Our results show that while the vast majority of genes have bounded expression levels consistent with stabilizing selection, most gene expression trajectories show significant evolutionary divergence in timing patterns. An organism's transcriptome may be able to acquire adaptive, functional changes in gene expression through changes in the timing patterns of dynamically-autonomous gene modules, potentially alleviating the negative pleiotropic effects associated with changes in regulatory interactions and changes in the TF expression levels.

3.3 Results

3.3.1 Genome-wide expression levels show much less variability than expected, but CDC-temporal expression patterns display broad divergence

We assayed transcriptome levels throughout the mitotic cell-division cycle (CDC) of 10 natural budding yeast lines, including 8 woodland and 1 laboratory strain of *S. cerevisiae* and 1 outgroup of *S. paradoxus*, in a comparative experimental design (see Materials and Methods and Figure 3.1). To calibrate gene expression variation across these lines with an expectation from mutation–drift, we also measured the transcriptomes of 23 mutation accumulation (MA) lines. Normalization and processing of our data yielded expression

levels for 6263 genes at 18 sampled CDC-timepoints for natural lines and unsynchronized expression for MA lines. To assess evolutionary variability in gene expression, we computed F-statistics for 4973 genes with significant mutational variance (*2*) (FDR < 0.25) as the ratio of per-generation natural to mutational variances within *S. cerevisiae* (d.f. 8 and 22). The genome-wide CDC median F-value is 1.56×10^{-4}, indicating that variation among natural strains is roughly 10^4-fold smaller than expected under mutation–drift equilibrium, which is 1.54×10^{-4} (*cf.* (*17*)). When tests are carried out for each gene at each timepoint (Figure 3.2A), 95.6% of genes exhibit stabilizing selection on expression level on average (FWER $< 10^{-5}$). With a maximum F-value of 0.23, not a single gene appears to have undergone positive selection for functional adaptation at any timepoint. The 9 natural *S. cerevisiae* strains in our study are estimated to have diverged between 3.02 and 4.19 thousand years ago (95% C.I.); therefore 94.4% to 96.4% of genes are under stabilizing selection. Moreover, the majority of genes (81.9%) exhibit expression trajectories consistent with complete stabilizing selection at every timepoint, while 742 genes (15.0%) exhibit low variability in at least half of the timepoints (partly neutral trajectories), and only 152 genes (3.1%) exhibit neutral variability in at least half of the timepoints (neutral trajectories) (Figure 3.2D). Thus only a small fraction of genes have predominantly neutral expression trajectories; no single trajectory appears to evolve completely neutrally. This static view our data is consistent with previous hypotheses that the expression levels of most genes are under strong stabilizing selection.

Temporal signatures of gene expression variation

This broad pattern of transcriptome-wide stabilizing selection could reflect an overall lack of temporal fluctuations in CDC-expression. To test this, we partitioned expression variation into relative contributions from strain and temporal effects using a linear mixed model

analysis. 3750 genes (59.9%) exhibit significant effects (FDR < 0.1 over all 6251×2 hypotheses): 2797 genes (46.6%) show significant strain variation (*i.e.* divergence), 2596 genes (43.3%) show significant temporal variation, and 1643 genes (26.2%) show both effects. Among these 1643 genes, strain and temporal variances do not strongly correlate ($R = 0.25$, Figure 3.3), and temporal effects contribute 10^4-fold more to overall expression variation than strain effects (genome-wide medians $\sigma^2_{time} = 9.54 \times 10^{-4}$ *vs.* $\sigma^2_{strain} = 7.43 \times 10^{-8}$).

In addition, clustering the entire expression data set shows a complex inter-relationship among strains and timepoints (Figure 3.4). Notably the *S. cerevisiae* laboratory strain exhibits the most divergent transcriptome in its dynamic profile, beyond the outgroup species *S. paradoxus*, despite having only 341 genes whose average expression levels differ significantly from woodland strains (t-test, FWER < 0.1). Thus, while expression levels show limited evolutionary variability, the dynamic pattern of expression displays broad between-strain and between-species divergence leading to a more complex view of transcriptome control and evolution.

Association of gene function with evolutionary forces

To relate evolutionary forces to yeast gene function, we computed the proportions of genes $Q_j(t)$ under stabilizing selection for 88 GO Slim terms j over time (Figure 3.5, Table 3.2). $Q_j(\cdot)$ profiles of most terms appear qualitatively similar, and 83 terms have average values $\overline{Q}_j$ greater than 0.94, indicating most processes are strongly affected by selection. The remaining 5 terms are the least affected: helicase activity (0.76), extracellular region (0.86), cell wall (0.91), cellular component (0.92), and pseudohyphal growth (0.93). Of these only cell wall and extracellular region genes are enriched among the 1643 temporally variable, divergent genes, while only cellular component genes are enriched among the 152 genes

with neutral trajectories (Fisher's Exact tests, FDR < 0.05). However there is significant overlap between the neutral and temporally variable, divergent gene sets ($P < 10^{-4}$), and cellular component encapsulates the others. This suggests that yeast cell surface structures are both temporally and evolutionarily variable. Alternatively a comparison of $\overline{Q}$ for 8 life-cycle related terms revealed that periodic, meiotic, and CDC-specific genes (in order, respectively) are the most neutral (Figure 3.2B); of these, only periodic genes are found in excess among the neutral genes (FDR < 0.05; Figure 3.2E, Table 3.3), while CDC-specific and ribosomal genes, as well as periodic genes are enriched among the temporally variable, divergent genes (FDR < 0.05). These results argue that periodic genes involved in cell structure, and more generally temporally variable genes, contribute the most to strain divergence. Moreover, the significant association of functional terms with patterns of time-dependent trajectories is consistent with the stabilizing selection operating on functional expression trajectories.

3.3.2 CDC regulatory architecture exhibits time-dependent changes in multi-dimensional complexity

To dissect the architecture of time-dependent control, we analyzed multivariate patterns of evolutionary expression covariation, including time-dependent multivariate patterns, using a variety of approaches. We first performed a canonical correlation analysis between all pairs of timepoints and found that expression between all pairs can be correlated perfectly ($R \approx 1.0$) via the primary canonical variables across strains (FWER < 0.05). The presence of such pervasive correlation across timepoints suggested that evolutionary divergence is not highly time-dependent. We next assessed the degrees of freedom of covariation and multivariate directions of evolution among strains using a latent factor mixed model analysis (LFA), principal component analysis (PCA), and singular value analysis (SVD); for

these the covariation at each timepoint was analyzed independently. Using LFA, we found that expression covariation can be explained by an average of 4.6 latent (*e.g.* genetic) factors with a range of 2 to 9 factors across the CDC. The temporal pattern of expression covariation suggests a degrees of freedom reduction of latent factors generally coincident with CDC-phase transitions G_1/S, G_2/M, and M/G_1 (Table 3.4), reflecting greater constraints on gene regulation, potentially through the influence of CDC checkpoints.

Conversely, PCA indicates that a slightly larger average of 6.1 factors is needed to explain 90% of variation at each timepoint (Figure 3.8A); if each strain's expression is time-averaged, a total of 5 PCA factors explain between-strain variation; if each timepoint's expression is strain-averaged, 10 factors explain between-timepoint variation (Figure 3.10). In contrast, MA line data, which are effectively time-averaged, reveal only 2 latent factors but 13 PCA factors. Thus, evolutionary divergence at any given timepoint not only contains more regulatory complexity than time-averaged data, but is also more restricted in degrees of freedom of expression evolution than expected by mutational input. Nevertheless, a total of 56 PCA factors are required for all timepoints and strains, revealing a much greater capacity for expression evolution when both CDC-temporal and strain covariation are taken into account. This implies a strong interaction between evolutionary strain divergence and the CDC, indicative of a time-dependent pattern of evolution.

To further dissect patterns of strain-CDC covariation we applied a SVD to expression data from all strains for each timepoint, obtaining 9 multivariate directions of evolution $U^r(t)$ for each of the 18 timepoints (*18*); we call these CDC-directions. We first reevaluated whether multivariate directions of evolution reveal a different, non-stabilizing pattern of selection compared to that of individual genes. F-statistics were computed using natural variance among woodland strains and neutral variance among MA lines using data projected onto each CDC-direction (Figure 3.11). Although the average F-value in the major

CDC-direction U^1 is 14.6-fold larger than the genome-wide average F-value (2.28×10^{-3} vs. 1.56×10^{-4}, $P = 1.5 \times 10^{-4}$), all F-values remain significantly consistent with stabilizing selection, including those calculated for minor CDC-directions (FWER < 0.05). Therefore, multivariate evolutionary patterns of transcriptome divergence are also consistent with stabilizing selection.

Both PCA and factor analyses revealed a complex time-dependent pattern of evolution, which suggests the presence of differential constraints on evolution as a function of CDC progression. We examined this by asking whether the CDC evolutionary covariance structure changes between different timepoints. First we computed angular distances between major CDC-directions for all timepoint pairs ($\angle U^1(s)\ U^1(t)$, Figures 3.12A and 3.12C). Adjacent timepoints as well as those in phase with the CDC appear more similar than other timepoints, indicating that changes in covariance structure are both gradual and cyclic. Despite these similarities, angles range from 19.4° to 88.9° and average 50.4°, suggesting most major CDC-directions are distinct. A random angles test failed to identify any significantly small angles, even with a lenient cutoff (FWER < 0.15). Similar testing of each of the 8 minor SVD directions (Figure 3.13) found only 8 small angles out of 1072 comparisons. Thus we observe significantly different patterns of evolutionary covariation throughout the CDC.

We also compared entire evolutionary covariance matrices by asking how many eigenvectors are shared between timepoints. Limited to 9 degrees of freedom (number of *S. cerevisiae* strains), we first projected expression data from each timepoint t onto that timepoint's top 9 CDC-directions $U^{1\text{–}9}(t)$ and subsequently computed a 9×9 covariance matrix for that timepoint. We compared covariance matrices between sequential timepoints using the common principle components (CPC) test (*19*). 15 of 17 comparisons reveal some significant difference in covariance structure (*i.e.* < 9 shared eigenvectors), but sequential

timepoints appear to share 5 eigenvectors on average (Table 3.5). The degrees of freedom restriction however only permits us to consider less than 50% of the total evolutionary covariation at each timepoint, and covariance structure estimates based on low dimensional data projections may be distorted. These preliminary results based on entire evolutionary covariance matrices also support the time-dependent nature of CDC evolutionary covariance structure.

Functional analysis and predictability of time-dependent covariance patterns

We identified the genes contributing most to the major CDC-directions and determined the functional terms enriched among the resulting top 5% of genes (Tables 3.6 and 3.7). The terms identified at each timepoint vary, but metabolic, periodic, and ribosomal terms (in order respectively; Table 3.8) are enriched in the major CDC-directions (FDR < 0.05). Few of the 152 genes with neutral trajectories (neutral genes; Figure 3.2D) are found among the top 5% ($P < 10^{-5}$), and only 1 neutral gene from each category ever appears among the top 5 ($P = 2.6 \times 10^{-3}$). Thus, neutral genes do not seem to drive strain-specific evolutionary divergence. TATA regulatory motifs also have been hypothesized to drive expression divergence via neutral drift (*17*). However, TATA-associated genes project onto major CDC-directions 4-fold less than genes lacking TATA motifs, which in fact are over-represented among the top 5% of genes ($P < 0.01$, Table 3.8). Overall the lack of evidence for neutral genes argues against drift as a major factor in strain diversification.

We also tested whether the within-species covariance patterns are predictive of directions of between-species divergence as might be expected for neutral species divergence (*20*). For each timepoint we calculated the angle between the major *S. cerevisiae* CDC-direction and the displacement vector of *S. paradoxus* expression, oriented within *S. cerevisiae* CDC-space (*e.g.* Figure 3.14). All angles exceed 45°, and no angle is signif-

icantly small. Thus, within-species covariation does not predict the direction of between species divergence. However, release from α-factor, S-phase, and the G_2/M transition have the smallest angles, suggesting that response to mating pheromone and DNA replication dynamics are more constrained in evolutionary variation.

Summary

To dissect the architecture of time-dependent control, we analyzed multivariate patterns of expression covariation among the *S. cerevisiae* lines, including time-dependent multivariate patterns. A much greater complexity of expression divergence is revealed when both CDC-temporal and strain covariation are taken into account. Significant changes in CDC-transcriptome covariance structure are observed across timepoints, in a manner that is both gradual and cyclic (Figures 3.12A, 3.12C, 3.13). Major directions of evolutionary covariation at different timepoints are significantly enriched for particular functional categories (Tables 3.6–3.8) and within-species covariance patterns do not predict directions of between-species divergence for our outgroup species *S. paradoxus*, as might be expected for neutral species divergence (*20*). Compared to mutational input, time-specific covariation exhibits greater regulatory complexity globally (mixed model factor analysis) and restricted dimensions of evolutionary expression covariation (principle components analysis) at every timepoint (Figures 3.8–3.10, Table 3.4). Thus, at each timepoint divergence is channeled towards putatively important functional directions, which themselves differ across timepoints. Preceding each CDC-phase transition (97, 218, 267, and $\approx$350 min.; except G_1/S), there is a large change in the major axis of CDC covariation (63, 152, 251, and 301 min.; Figure 3.12B), immediately followed by a peak of expression variability (87, 176, 260, and 345 min.; Figure 3.11) that coincides with a drop in overall covariation complexity (176, 260, 345 min.; Table 3.4). The average variability peak is both greater

than that at all other timepoints ($P = 0.018$) and 19.1-fold greater than the genome-wide average ($P = 0.006$). This temporal fluctuation in variability might reflect multi-genic (pleiotropic) effects being directed to varying dimensions and directions of gene expression through a regulatory architecture which changes across CDC-phases.

3.3.3 Divergence in coregulatory patterns is found at all scales of transcriptome organization

To evaluate strain divergence in temporal coexpression patterns, we computed a 6082×6082 gene coexpression matrix for each strain using the correlation across CDC-timepoints and then calculated matrix correlation coefficients between pairs of strains (Figure 3.15A). Due to the extreme size of the matrices, all comparisons yield significant concordance in coexpression patterns (FDR < 0.01), but the degree of concordance is low (avg. $R = 0.11$) indicating that most strains lack strong similarity in CDC-coexpression. *S. paradoxus* has the lowest coexpression correlation with other strains (avg. $R = 0.047$) while *S. cerevisiae* strains YPS3137 and YPS2073 also have low correlations (0.055 and 0.068). Restricting these coexpression matrices to 270 transcriptional regulatory genes does not strengthen this pattern of weak association (Figure 3.15B).

Controls on CDC-coexpression divergence

To assess the magnitude of observed levels of CDC-coexpression divergence across strains, we computed matrix correlation control statistics using technical and biological replicate control data. Comparison of a pair of dye-swap technical replicates of the laboratory yeast strain (independent of our data (*21*)), yields a matrix similarity of $R = 0.58$, compared to the average similarity of 0.114 shown in Figure 3.15A. As a biological control, we

deinterlaced time-series data of one woodland strain (YPS2055) into 2 subsets of odd and even timepoints, and computed correlations between them, obtaining $R = 0.18$. Although this biological control shows a much lower correlation than the technical control, only 9 timepoints were used to calculate the genome-wide correlation matrices (*vs.* 25), and the paired expression levels were on average measured 19 min. apart (*vs.* 0 min. for the technical control).

We obtained additional positive controls by comparing our CDC-expression data for strain YPS183 to data generated through a parametric resampling procedure, where Gaussian time-sampling noise was added to YPS183 expression data at each timepoint, using our estimate of 4.17 as the sampling variance around the mean of each true timepoint (see Materials and Methods). We generated 10 resampled YPS183 matrices and compared them to the actual YPS183 matrix, yielding an average matrix correlation of 0.778, with a minimum value of 0.761 and a maximum value of 0.796 ($P < 0.001$). To control for error due to biological replication, we also generated 10 matrices where, in addition to time sampling noise, Gaussian measurement noise was added to every data point (expression level of gene g at timepoint t) using mean y_{gt} and variance 0.307 (corresponding to the square of our estimated biological replicate error 0.554). Here the average matrix correlation is 0.371, with a minimum of 0.361 and a maximum of 0.380 ($P < 0.001$). This average (matrix similarity among simulated biological replicates of the same strain) is more than 3 times larger than the average matrix similarity of 0.12 between strains.

Since the large size of these correlation matrices contributes to the significance of each Mantel test, we also computed correlations for 100 resampled matrices using only the expression data for 270 transcription regulators (as in Figure 3.15B). Adding only sampling error, matrix correlations average 0.771 ($P < 0.001$) with a minimum of 0.721 and a maximum of 0.800. Adding both sampling error and measurement error, correlations average

0.304, with a minimum of 0.265 and a maximum of 0.348 ($P < 0.001$).

These results demonstrate that errors either in measurement or in culture sampling cannot be attributed to the low observed matrix correlations between strains. Rather these results suggest a pattern of significant CDC coexpression divergence between strains, both within and between species.

Time and scale-dependent evolutionary divergence of coexpression structure

Given the temporal nature of CDC covariation, coexpression divergence might specifically accumulate in particular CDC-phases. To assess coexpression concordance in a time-specific manner, we grouped each strain's expression data into 3 overlapping CDC-phase groups (first, middle, and last 9 timepoints) and again assessed correlation matrix similarity between strains. While divergence is seen both across phase-subsets per strain and across strains per phase-subset (overall $\overline{\mathrm{R}} = 0.075$), coexpression matrices clearly cluster by strain (Figures 3.16A, 3.16B). However the cluster relationships among strains are unique to each phase-subset (Figures 3.16C, 3.17), indicating evolutionary coexpression divergence is again time-dependent.

Evolutionary divergence of the yeast coexpression structure may also occur at any or all scales of transcriptome organization. We characterized transcriptome coexpression structure at various scales of organization by defining a coexpression module for every gene as its k most correlated genes within each strain, for 6 values of k: 25, 100, 500, 880, 1344, and 2500 genes. We then assessed the modular similarity of gene coexpression by computing the overlap of each gene's k-modules between all pairs of strains. The degree of shared coexpression structure between a pair of strains is the proportion of genes with significant k-module overlap ($P < 1/250$; Table 3.9). Less than two-thirds of the genes exhibit significant overlap at any scale (from 25% at $k = 25$ to 65% at $k = 2500$, averaged

over all strain pairs), suggesting that signatures of shared temporal coexpression cannot be identified for a large portion of the genome. While there is consistently greater average overlap of significant genes than expected by chance using binomial sampling, the amount of excess is generally low (averaging 8.24% with min= 4.39% at $k = 25$, max= 10.03% at $k = 880$; Figure 3.18, Table 3.9). Thus, similar to the matrix correlation results, the overall pattern of modular coexpression shows low concordance between strains regardless of scale.

To determine whether relationships of shared coexpression structure between strains change across organizational scales (*i.e.*, values of k), we computed hierarchical clusterings of the 10×10 matrices of average module overlap between strains for each k (Figure 3.19). The overlap relationships between strains change at different scales, with lower overlap at smaller scales, suggesting that transcriptome coexpression structure evolves more rapidly for genes that are more tightly coexpressed within a genome. For a subset of strains, the overall module overlap relationships stay constant at most scales. However a few strains, notably YPS3137 and YPS2073, show considerable variation in module overlap across scales, suggesting that these strains differ in their coregulatory relationships at all scales of coexpression.

3.3.4 A hypothesis for modular timing control

The results above indicate surprisingly large divergence in expression dynamics, suggesting potentially adaptive evolution of expression *dynamics* despite stabilizing selection on expression *levels*. Changes in dynamics imply changes in the timing patterns of genome-wide gene expression. If the control of timing patterns involves a global cascade of regulation, any changes in control could cause broad negative pleiotropic effects throughout the CDC (*7*). Alternatively, if the architecture of genome-wide gene regulation is orga-

nized into discrete timing modules, then the superposition of different timing patterns could combinatorially generate the regulatory complexity required for transcriptome-level timing control while minimizing negative pleiotropic effects.

3.3.5 Heterochrony explains evolution of expression dynamics

We evaluated the hypothesis of modular timing control by identifying genes that share patterns of heterochronic evolution (*22, 23*), which may be used to delineate dissociable units of structure and function (*24, 12*). First, we tested for the presence of heterochronic evolution by asking whether a time transformation (*i.e.,* heterochrony) model significantly explains a gene's divergence in CDC-expression between two genomes (Figure 3.20A). On average, our heterochrony model explains 61% of between-genome transcriptome variation (Figure 3.20B). We computed a likelihood-ratio statistic for every gene, comparing the fit of the heterochrony model to the fit of a time-independent model. Between 64% and 96% of genes are significant for any between-genome comparison (d.f. 3 and 14, FDR < 0.05; Figure 3.20C), indicating a broad pattern of heterochronic evolution. Each gene exhibits significant fit to the heterochrony model for an average of 33.1 of the 45 pairwise comparisons (Figure 3.20D). These results suggest that the major mode of transcriptome evolution entails changes in timing (heterochrony) rather than changes in levels (heterometry) of expression.

Functional characterization of heterochronic and non-heterochronic genes

We kept 4998 genes showing consistent support for heterochrony (genes with $\geq 2/3$ sig. comparisons; Figure 3.20E). As expected, these heterochronic genes tend to exhibit dynamical fluctuations in expression level throughout the CDC (85.8%, $P < 10^{-10}$) (Sec-

tion 3.3.1). These genes are enriched with a variety of GO functions, including (in order of decreasing significance) ribosome biogenesis and assembly, cytoplasm, nucleolus, oxidoreductase activity, transferase activity, mitochondrion, carbohydrate metabolic process, generation of precursor metabolites and energy, hydrolase activity, transporter activity, cellular bud, amino acid and derivative metabolic process, other, electron transport, endoplasmic reticulum, vacuole, cell wall, lipid metabolic process, plasma membrane, translation, and cofactor metabolic process. In addition there is an association between genes that are heterochronic and genes that exhibit neutral expression trajectories ($P = 0.024$; Figure 3.2D). Notably an excess number of neutral, heterochronic genes have a GO function of helicase activity (FDR < 0.05), reflecting the finding that the set of all genes with helicase activity evolve under the least amount of stabilizing selection on expression levels (Section 3.3.1). Conversely we defined the complement of the set of heterochronic genes as those exhibiting time-independent evolution. 369 of these 1084 non-heterochronic genes exhibit dynamical fluctuations in the CDC, constituting a class of genes with significant time-dependent gene expression patterns that may be evolutionarily conserved. These genes associate with GO functions corresponding to response to chemical stimulus and conjugation (FDR < 0.05).

3.3.6 Shared patterns of heterochrony reveal modular timing changes

Timing modules may be revealed by patterns of heterochronic evolution (*22, 23*). We characterized patterns of heterochronic evolution (via timing curves) for each gene between strains using a time transformation (*i.e.* heterochrony) model, where the estimated model parameters define a timing curve (or timing pattern change) for a gene between two strains. If changes in expression timing are functionally meaningful, genes sharing similar timing curves across many strains may belong in the same functional timing module.

Clustering of gene expression timing pattern changes

To identify groups of related genes, we compared timing curves between the 4998 genes showing consistent support for heterochrony. To avoid any bias in our estimation procedure, we obtained an error-bounded ensemble of timing curves for every gene in each strain comparison. Each ensemble constitutes the set of timing curves which characterize the evolutionary time-domain transformation between expression trajectories equally well (within 95% of the optimal timing curve). The ensemble size for most genes is small, averaging 26.8 timing curves (Figure 3.22A). Thus, we defined the relationship between any pair of genes by the distance between each gene's 95%-equivalence timing curve ensemble. The distance between any 2 timing curves was computed as their average root mean squared error (RMSE), thus integrating the differences between curves across the entire CDC. Since our time-domain parameter estimation procedure may be somewhat coarse-grained, we then defined the distance between 2 genes as the minimum RMSE distance over all pairs of timing curves between the ensembles of the 2 genes. In this way 2 genes are similar if their timing curves are concordant across the entire CDC (Figure 3.23). Using this metric, we computed a timing pattern distance matrix M_i for each of the 45 strain comparisons (*e.g.,* Figure 3.24).

To classify genes into timing pattern groups, we first used metric multidimensional scaling (MDS) to coordinatize the pairwise distance relationships, such that the Euclidean distances between genes in this space match those in M_i (typically around 1500 out of a possible 4997 dimensions captured $> 100\%$ of the variation in each M_i). Using these coordinate matrices, we then clustered genes with a k-means procedure. (*N.b.,* hierarchical clustering of M_i directly using Ward's criterion generated similar clusters; data not shown). A range of k values yielded a clear partitioning of the data for each of the 45 strain compar-

isons, and visualization of the mean timing patterns for each cluster shows they are distinct (Figure 3.25). However, genes that are functionally related should show similar distances across all 45 strain comparisons. We computed a summarized distance matrix M by taking the element-wise averages across M_i (that is, averaging over each gene pair's 45 distance values). Distance values might reflect slight variations of each relationship, so averaging over all 45 comparisons should estimate the true distance relationships. A gap statistic analysis using k-means showed significant support for 7 clusters in the summary matrix M (Figure 3.26A). This suggests a CDC timing control architecture composed of a set of modular timelines of gene expression events.

Robustness of timing pattern clusters

To assess the support for gene–module classification, we bootstrapped the summary matrix M by randomly resampling (with replacement) 45 new M_i and computing a resampled summary distance matrix. We then clustered genes using this new matrix and computed concordance statistics as the average fraction of genes that are identical between real and resampled clusters. We find an overall concordance of 22.6% over 100 bootstrap resamples, with little variation across clusters (Figure 3.26B). We also performed a jackknife analysis to assess cluster sensitivity to the particular spatial distribution of points. On average the clusters show 30.7% concordance. These results suggest that many genes may not associate clearly with individual timing modules. However, linear discriminant analysis clearly distinguishes among 7 clusters for each of the 45 distance matrices (Figure 3.27). This potential discrepancy between individual and summary clusterings might be due to the coarse grained nature of our parameter estimation or the low sampling frequency of expression data, such that averaging over distances actually diffuses the pairwise relationships. However, another possibility is that the CDC timing control architecture contains

distinct timing modules described by core sets of genes which consistently group within particular modules.

Identification of modular heterochronic interactions

To identify the genes significantly associated with each module, we performed a pairwise analysis by counting the number of between-genome comparisons (out of $\binom{10}{2} = 45$) in which two heterochronic genes share timing changes. Overall, 5393 significant interactions connect 3715 genes (binomial, $P < 10^{-4}$; Figure 3.28), comprising a heterochronic network of interactions within and between modules. To identify the genes significantly interacting within each module, we mapped the heterochronic network to the 7 modules obtained from M (Figure 3.29). 2546 interactions (47.2%) connect 2323 genes within individual modules (shared heterochrony), and 2847 interactions (52.8%) connect 1392 genes between modules. Compared to all possible interactions within and between modules, shared heterochronic interactions are found 5.6-fold more often than significant interactions between modules ($P < 10^{-10}$), indicating that genes clustering together exhibit similar timing changes.

We focused on the 1828 genes whose interactions indicate strongly shared heterochrony (Figure 3.29); these genes comprise the core of each module. Linear discriminant analysis is able to distinguish between the modules defined by these genes (Figure 3.31). Genes interacting within a module tend to share functional ontology terms, on average sharing 95% of possible life-cycle terms ($P < 10^{-7}$) and 23% of possible GO-slim terms ($P < 10^{-19}$). GO enrichment analysis using genes with shared heterochrony reveals the functional identities of each module (Table 3.11); similar analysis of the set of between-module genes did not reveal any significantly enriched terms (FDR < 0.1). Thus, timing modules contain core sets of closely associated genes, supporting a CDC timing control architecture

involving distinct timing modules.

Significant modularity in expression timing patterns suggests that each timing module might undergo dynamically-autonomous timing pattern evolution. We assessed timing pattern variability among modules using the 1828 within-module genes by pooling optimal timing curves over the 45 strain comparisons for the genes in each timing module and testing for significantly different mean timing patterns using an analysis of variance. Modular timing patterns appear to differ significantly (ANOVA, $P < 10^{-10}$), suggesting that each timing module might undergo heterochronic evolution in a dynamically-autonomous manner. Taken together, these results suggest that the CDC timing control architecture is comprised of a core of dynamically-autonomous timing modules, involving nearly 30% of genes genome-wide, combined with a layer of interactions across modules, which could perhaps serve to coordinate or synchronize expression timing globally.

Variability of modular expression timelines

The 7 timing modules show significantly different timing patterns and associate significantly with both functional ontology terms and regulatory factors. Thus each timing module could represent a distinct component of temporal development responsible for executing a particular timing pattern. If this were true, control of the timing of gene expression in each module might be coherent. That is, variation in timing patterns among the genes within an individual module may be lower than expected by chance. Although the k-means clustering procedure (used to identify timing modules) in part tries to minimize this timing pattern variance in each module (*i.e.* within cluster error), this minimization is relative to the distribution of timing patterns seen in only one particular strain comparison. However, these variance values may not differ from those expected in the space of all possible timing patterns. We assessed the significance of modular timeline variability by comparing the

observed variance in timing patterns in a timing module to a random distribution of cluster variances, where each random variance results from grouping random timing patterns (see Materials and Methods). We applied this test to the timing curves estimated for the 36 comparisons between *S. cerevisiae* strains and assessed which modules show significant variability using a false discovery rate of 0.001, per comparison. 34.1% of each module's comparisons are significant on average (12.3/36), and the proportion of significant comparisons varies across modules from 0.69 (module 1) to 0.06 (modules 2 and 5) (Figure 3.32, Table 3.12). Although all 7 modules show significant variability in some comparisons, modules 1, 3, 5, and 6 together have an average significance of 53%, while modules 2, 4, and 7 show an average significance of 8.3%, indicating that only some modules consistently show significant variability. Modules 3 and 6 show significantly high comparisons (3 each), so the vast majority of significant comparisons indicate low variability. We also assessed variability for comparisons between species. Only 3 of the 7 timing modules showed any significant comparisons and the number of comparisons is low (5, 3, and 2 for modules 1, 5, and 6, respectively). Thus, modular timeline variability appears lower within species than between species. These results suggest that for some timing modules modular timing pattern variability tends to be low.

It is possible that low patterns of variability could result from an overall similarity of timing patterns for genes across different timing modules. Although an ANOVA test found that timing modules have significantly different timing patterns (Section 3.3.6), we also explicitly tested whether low modular timing pattern variance results from the particular grouping of genes in each module, by shuffling the membership of genes in timing modules and computing the variance within each shuffled module. We then compared the total within-module timing pattern variance to a distribution of 100 shuffled-module variances for each strain comparison. Observed timing pattern variance is significantly low in all 45

strain comparisons (FDR < 0.05). Thus the pattern of low modular timing pattern variance is related to the specific grouping of genes within modules.

The idea of low modular timeline variability would also be supported if genes that are heterochronic but not module-specific show a different variability pattern. We assessed whether the remaining set of 1887 between-module genes exhibits significant timeline variability. These genes show a comparable fraction of within-species comparisons with low variability (47.2%), but more comparisons with high variability (6, compared to 3 for modules 3 and 6). In general heterochronic genes show limited timeline variability for many strain comparisons, but genes within timing modules tend not to show high variability. Thus, levels of timeline variability may be module specific and overall tend to be low, consistent with the idea that timing modules execute coherent gene expression timelines.

Evidence for canalization of modular expression timelines

Genetic variation expressed through a timing control architecture allowing pleiotropic effects could precipitate dramatic changes in expression timelines within modules, potentially leading to chaotic expression dynamics (*25, 26*) and the failure to complete a round of cell division. However, modular gene expression timelines tend to be similar throughout the entire CDC and appear to play functional roles in CDC progression, suggesting that changes in expression timing within a module could have deleterious consequences. In this case, it is possible that variability of modular expression timelines may be limited by a form of negative selection, potentially canalizing selection (*27,28*), which could reinforce the coherence of modular expression timelines as integrated developmental processes. If natural selection influences variability of modular expression timelines, then the natural evolutionary timeline variability of each gene, as well as the overall evolutionary variability of each module, should be significantly limited within species.

We first assessed evolutionary timeline variability for each of the 4998 heterochronic genes by computing each gene's natural evolutionary timing pattern variance across the 36 within-*S. cerevisiae* comparisons and comparing each variance value to a random variance distribution (see Materials and Methods). 1244 genes (24.9%) show significantly low variance, and 117 genes show significantly high variance (2.3%) (FDR < 0.05). Genes with low variance associate with CDC, ribosomal, and periodic life-cycle terms (FDR < 0.05), while genes with high variance show no significant life-cycle term associations. Intriguingly, 526 of the 1244 genes with low evolutionary variance match within-module genes (42.3%, $P < 10^{-5}$), while only 28 of the 117 genes with high variance match (23.9%, $P = 0.006$). This indicates a specific enrichment of genes with low evolutionary timeline variability in the set of within-module genes ($P = 0.0002$) and suggests a relationship between coherence and evolutionary variability of modular expression timelines.

We next assessed modular evolutionary timeline variability by asking whether the mean evolutionary timing pattern variance over the genes in a module differs from the mean evolutionary variances of random groups of heterochronic genes. We created a distribution of 10,000 random variances by computing the mean evolutionary timing pattern variance of a set of genes with the same size as a true module, but chosen by randomly sampling from the set of 4998 heterochronic genes. Significance of this test indicates that the particular membership of genes in a timing module confers low evolutionary variability. 5 of the 7 modules display significant variability (FDR < 0.05): modules 1, 3, and 5 show low variability, while modules 2 and 4 show high variability. Variability in modules 6 and 7 is low but not significant. This suggests that modules 1, 3, and 5 might experience a form of negative selection on expression timelines, while modules 2 and 4 might experience selection pressure to maintain different timelines, *i.e.* timeline plasticity. The strength of modular evolutionary variability correlates strongly with the coherence of modular expres-

sion timelines (number of strain comparisons showing significantly low modular timeline variability) (Spearman's $r = -0.94, P = 0.0009$). Such a relationship between evolutionary and developmental timeline variability would easily be explained if each gene in each timing module simply had the same timing pattern across all of the strain comparisons, thereby producing low evolutionary timing pattern variance values. However, modules 2 and 4 do have significantly high variability, so there must be at least some genes which do change substantially across strain comparisons. Also, modules with high evolutionary variability may not be stable by measures of module stability, such as the computed bootstrap and jackknife concordance statistics (Section 3.3.6). So if modules 2 and 4 (having high mean evolutionary variability) are also the least stable, this would suggest that the relationship between modular evolutionary variability and modular timeline variability is false. While module 4 does show the lowest bootstrap concordance (0.188), module 2 shows highest bootstrap concordance (0.273) (Figure 3.26B). Also, jackknife concordance shows that modules 4 and 2 are the 3^{rd} and 4^{th} most stable modules (data not shown). There is also little correlation between evolutionary variability and bootstrap (Spearman's $r = 0.1$) or jackknife (Spearman's $r = 0.0$) concordance across the 7 timing modules. Thus, the pattern of modular evolutionary variability does not correlate with the pattern of module stability, consistent with a biologically meaningful relationship between modular evolutionary timeline variability and modular developmental timeline variability. This potentially suggests that the reason modules show low developmental timeline variability is because of natural selection for modular timeline coherence.

Evidence for inter-module coordination of expression timelines

Despite potential selection for modular timeline coherence, timelines appear dynamically-autonomous and may evolve independently across modules. If modules were free to evolve

independently, large relative changes in the execution of expression timelines between modules could disrupt the functional integration of the CDC (*29*) or possibly result in unstable expression dynamics. If relative coherence of modular expression timelines is an important aspect of CDC progression, then expression timelines may be coordinated or synchronized across timing modules. Signatures of inter-module coordination might be revealed as bottlenecks in timing pattern variability among modules at specific CDC timepoints. To evaluate this, we computed variance in timing patterns among all 1828 within-module genes at several timepoints across the CDC. We then scaled each variance value by the median of a random variance distribution for that timepoint. A plot of these comparisons across the CDC reveals a time-dependent pattern (Figure 3.33) with local minima in global timing variability closely following estimated CDC checkpoints. There is lower variability than expected overall, and variability tends to increase following one checkpoint and decrease preceding the next. These major features are also seen in the curve of multivariate expression-level variability along major CDC-directions (Figure 3.14). One unique aspect of this timing pattern variability curve is that variability generally remains low until late G_2-phase, where it quickly rises through the end of mitosis then falls dramatically to levels seen in early G_1-phase. This suggests a possible loss of timing control or relaxed selection pressure on inter-modular coherence at the end of the CDC, preceding the re-synchronization of timing modules at the beginning of a new cycle. Overall, these results suggest that evolution of the scheduling and coordination of gene expression timelines, vis-à-vis the CDC timing control architecture, may represent an important mode of evolution.

Summary

Clustering genes by their timing changes revealed 7 significant modules (Figures 3.26A), consistent with the hypothesis of modular timing control. To identify the genes significantly associated with each module, we performed a pairwise analysis by counting the number of between-genome comparisons in which 2 heterochronic genes share timing changes. Overall, 5393 significant interactions connect 3715 genes (binomial, $P < 10^{-4}$; Figure 3.28); 47.2% of significant interactions connect genes within individual modules (shared heterochrony). Compared to all possible interactions, heterochronic interactions are found 5.6-fold more often than significant interactions between modules ($P < 10^{-10}$), indicating that genes sharing timing pattern changes tend to cluster together. We focused on the 1828 genes whose interactions are mostly heterochronic (Figure 3.29); these genes comprise the core of each module. Genes sharing heterochronic interactions share functional ontology terms, on average sharing 95% of possible life-cycle terms ($P < 10^{-7}$) and 23% of possible GO-slim terms ($P < 10^{-19}$). Overall variability in the modular patterns of timing change among these genes and among strains is lower than expected, suggesting potential canalization of timing modules as distinct temporal units, or event timelines, of the CDC (Section 3.3.6). In support of this, modules exhibit significantly different patterns of timing change (ANOVA, $P < 10^{-10}$; Figure 3.31), suggesting that event timelines are dynamically-autonomous and can evolve differentially. At the same time, global variability among all patterns of timing change is lower than expected and time-dependent, suggesting coordination and periodic synchronization of event timelines (Section 3.3.6). These results indicate significant modular organization in the timing patterns of genome-wide gene expression.

3.3.7 A modular, pleiotropic regulatory architecture explains CDC transcriptome divergence

Since genes in timing modules tend to share TFs, it is possible that some TFs might specifically control expression timing across the CDC of the genes in each timing module. We tested a set of 169 TFs (the subset of 204 TFs from (*5*) which interact with the 1828 within-module genes) for association with each timing module. For each TF and each module we made a 2×2 contingency table by counting the interactions between one TF and the genes within one module, the TF's interactions with genes in all other modules, the interactions of all other TFs to one module, and their interactions to all other modules. We identified 25 TFs showing module-specific associations (Fisher's Exact test, $P < 0.05$; Table 3.13); on average 3.6 TFs associate with each module. Heterochronic patterns of evolution are prevalent among these 25 TFs, with an average of 72% of significant heterochrony model tests between strain comparisons. As a class, these TFs also show elevated distortion compared to expectation from all heterochronic genes (76^{th} percentile) and from all TFs (76^{th} percentile)—the expected distribution consisted of the median distortion of 25 genes randomly sampled without replacement 10,000 times from the set of all heterochronic genes or all TFs, respectively. There do not appear to be significant differences in the distortion of TFs between modules (ANOVA, $P = 0.2$). In addition, the TFs exhibiting the most distortion in each timing module show significantly elevated distortion levels compared to all heterochronic genes and all regulatory factors (2-sample t-tests, P < 0.05), although only 1 of these TFs (CIN5) is among the top 50 of all heterochronic genes genome-wide (rank-46; see Table 3.10). Thus, regulatory factors undergoing heterochronic evolution may specifically control gene expression timing patterns in a module-specific manner (Figure 3.30A). Moreover, the number of module-specific factors may be small.

Conversely, a similar TF association test using the set of 1887 between-module genes revealed a single significant TF, ZAP1, which is a homeostatic regulator of cellular zinc levels (*30*). While the number of significant strain comparisons is high for ZAP1 (82.2%), its degree of distortion is significantly lower compared to module-specific TFs ($P < 0.001$; Table 3.13). Although we have only 1 statistic for comparison, this reinforces the notion that module-specific TFs may preferentially harbor elevated levels of heterochronic evolution.

This analysis of a combined heterochronic and regulatory interaction network suggests that timing pattern changes in those TFs that are highly associated with heterochronic genes within timing modules could predominantly contribute to the CDC-transcriptome divergence of expression dynamics. Thus, the widespread nature of CDC-transcriptome divergence might be explained by pleiotropic changes in the control of timing modules, perhaps mediated by genetic variation in relatively few regulators, which can induce timing pattern changes throughout functional gene modules. Furthermore, TFs themselves participate in 244 heterochronic interactions (*e.g.* Figure 3.30B). Since these TFs also serve as regulators of other interactions, timing changes originating upstream could propagate throughout timing modules via these TF intermediates.

Combined functional analysis of timing modules

Integration of results from GO analysis, TF–module association, and modular timeline variability analysis may reveal potential functional roles of each timing module. Of course, a cell is capable of performing a large variety of functions, and so the relatively few timing modules identified likely reflect a composition of several functions. Thus, the life-cycle terms, GO terms, and TFs associated with each timing module likely reflect some of the most prevalent aspects of each timing module. The 4 modules exhibiting the least modular

timeline variability (modules 1, 3, 5, and 6) show functional enrichment with metabolic, periodic, and ribosomal genes, while the 3 most variable modules (2, 4, and 7) show no functional enrichment, suggesting that the coherence of modular timelines could have a functional basis. Moreover, all timing modules show significant associations with transcription factors. Module 1 appears to function early in the CDC, that is preceding S-phase. RAP1 serves as an environmentally sensitive activator and repressor of transcription which has general chromatin remodeling abilities. Along with UME6, it is also involved in regulation of meiosis. FHL1 is an environmentally sensitive transcriptional activator involved in the expression of ribosomal protein genes. Together, these TFs may contribute to control of yeast reproductive mode. In addition, MSN4 is sensitive to various cellular stresses, notably glucose starvation, while CIN5 mediates pleiotropic drug resistance and salt tolerance. These TFs may help regulate intracellular environment during growth. Module 6 is also associated with FHL1 and ribosomal genes, but may also involve timing control of galactose metabolism (GAL3). Module 3, on the other hand, involves more periodic genes, and appears to involve timing control of late S-phase-specific genes (NDD1) and G_2/M genes (FKH2), possibly involved in cell-type-specific pheromone response (FKH2 and MCM1) and respiratory gene expression (HAP2 and HAP3). The roles of module 5 are unclear, but they may involve ribosomal and metabolic genes. As these modules generally associate with cell growth and reproductive life-history, tighter control of variability in their timelines may be important for reproductive fitness. In contrast, modules 2, 4, and 7 appear to associate predominantly with TFs involved in various stress responses, such as the unfolded protein response (HAC1), toxic damage response (MIG3), and multi-drug resistance (YRR1). They also associate with control of nucleosome assembly and positioning (HIR1, HIR3, CBF1), iron utilization and homeostasis (RCS1), and phosphate metabolism (PHO2). Module 7's association with SWI5 suggests function late in the CDC during M

and G_1-phases. Elevated levels of timeline variability in these modules could reflect a need for more flexible timeline control in responding to a variety of environmental perturbations or simply control over a more diverse set of biological functions. Overall, these results suggest that all timing modules are functionally relevant for CDC progression and may be responsible for execution of several different biological functions.

Summary

While the prevalence of heterochrony is consistent with broad changes in gene regulation, modularity in the patterns of heterochrony suggests that regulatory architecture itself may effectively constrain evolutionary variation into distinct channels of phenotypic expression. In this way, widespread divergence in transcriptome dynamics may be explained by changes in the expression of module-specific transcription factors, rather than changes in regulatory interactions per se. We used the 1828 genes with strongly shared heterochrony to test whether sharing heterochrony patterns implies common transcription factor transregulation. Genes sharing heterochronic interactions do share more TFs than expected ($P < 10^{-100}$) and associate with TFs more strongly than pairs of genes without significantly shared heterochrony ($P < 10^{-10}$). We then identified 25 TFs that associate specifically with each module (averaging 3.6 TFs per module). Module-specific TFs themselves exhibit significant patterns of heterochrony (Table 3.13), and at least one of the TFs in every module shows significantly large timing changes compared to all heterochronic genes or all regulatory factors ($P < 0.05$; Section 3.3.7). Thus, heterochronic changes in the expression of module-specific TFs may be a primary cause of divergence in transcriptome dynamics.

3.4 Discussion

Since changes in the regulation of gene expression are expected to have broad negative fitness effects, how do organisms achieve adaptive functional divergence of genome-wide gene expression? We hypothesize that adaptive divergence of genome-wide gene expression may be driven by heterochronic changes in the temporal expression patterns of relatively few transcription factors, inducing broad heterochronic changes in dynamically-autonomous gene modules. This hypothesis argues that the underlying adaptive genetic mutations should be found in loci affecting the expression of module-specific TFs, rather than loci affecting regulatory (TF–DNA) interactions. To assess this hypothesis we investigated the evolution of genome-wide gene expression as a dynamical system, using the first large-scale collection of comparative time-series transcriptome data covering the mitotic cell-division cycle of 10 lines of natural woodland budding yeast. We focused on evaluating the extent of potentially functional timing changes, the mode of time-dependent transcriptome evolution, and the architecture of timing control in the yeast cell-cycle transcriptome.

Our results show that while the evolution of genome-wide gene expression levels is consistent with strong stabilizing selection at each timepoint of the yeast cell-cycle (when compared to independent expression measurements from yeast mutation accumulation lines), a large fraction of genes show significant divergence in their dynamical patterns of expression. The time-specific pattern of transcriptome covariation among strains reveals gradual and cyclic changes in the preferred directions of divergence, and both the amount of covariation along these preferred directions and complexity in the pattern of covariation show distinctly time-dependent changes. That is to say, evolution in the yeast cell-cycle transcriptome is highly time-dependent, suggesting the broad potential for adaptive changes in

expression dynamics despite strong stabilizing selection on mean expression levels.

Since genome-wide changes in temporal expression patterns suggest broad changes in gene coregulation, we characterized the extent to which the structure of gene coregulation is shared across strains, finding significant divergence in the pattern of genome-wide temporal expression, across the entire CDC and in a time-dependent manner, as well as divergence in the pattern of temporal coexpression at all scales of structural organization. Of course divergence in temporal coexpression does not guarantee divergence in coregulation; two genes may be coregulated yet exhibit distinct temporal expression trajectories (or vice-versa). Thus we evaluated the possibility of heterochronic evolution, relating genes by shared changes in expression, rather than by similarity in expression levels (*i.e.* coexpression). The majority of genes show timing changes across strains consistent with heterochrony, suggesting that the major mode of transcriptome evolution involves changes in timing (heterochrony) rather than changes in levels (heterometry) of expression.

The prevalence of heterochrony implies broad changes in gene regulation, which are expected to have deleterious consequences in natural populations, such as our yeast strains, given a cascading regulatory architecture. However negative pleiotropic effects could be minimized by organizing regulatory architecture into different timing modules, each with a distinct timeline for gene expression. Analysis of the genome-wide patterns of heterochrony revealed significant modularity in yeast regulatory architecture, which is potentially established by relatively few module-specific transcription factors. Thus widespread divergence in transcriptome dynamics is best explained by heterochronic changes in the temporal expression patterns of module-specific transcription factors inducing broad modular heterochronic changes, rather than changes in regulatory interactions, per se. An organism's transcriptome may be able to acquire adaptive, functional changes in gene expression through changes in the timing patterns, or timelines, of dynamically-autonomous gene

modules, potentially alleviating the negative pleiotropic effects associated with changes in regulatory interactions and changes in the expression of globally-pleiotropic TFs while allowing evolution of complex patterns of regulation through the combinatorial superposition of timelines.

Our data suggest a new view of molecular cell processes as a collection of dynamically-autonomous event timelines whose modularity allows the adaptive divergence of gene regulation, while alleviating system-wide negative effects associated with regulatory change. Control of gene expression may utilize a general architecture comprised of multiple discrete event timelines which serve as an elemental, or basis set of timing patterns. Interactions among module-specific transcription factors may determine these event timelines, and the superposition of different timelines may be used to generate combinatorial complexity in regulatory patterns. In this way, the architecture of genome-wide gene regulation in yeast may evolve via modular changes in timing control, perhaps mediated by pleiotropic genetic variation in relatively few regulatory loci. This modular dynamical architecture may facilitate the generation of complex regulatory variation for evolutionary adaptation via changes in the scheduling and coordination of discrete event timelines, while buffering expression level changes in individual genes.

3.5 Materials and Methods

Summary of data collection and processing

Using 2-channel spotted oligo glass microarrays, the mRNA expression levels (intensities) of 6360 protein-coding genes were measured in 10 haploid yeast strains—8 woodland strains of *S. cerevisiae*, 1 derivative of laboratory strain S288c, and 1 woodland strain of sister species *S. paradoxus*—grown in synthetic dextrose (SD) minimal medium at 18°C (225

rpm). Cultures of heterothallic MATa derivatives of the natural homothallic diploid isolates were synchronized using α-factor mating pheromone, released from arrest before the G_1/S cell-division cycle (CDC) checkpoint, and sampled for RNA at 18 discrete timepoints over $\approx$1.3 mitotic CDCs (with 19 min. intervals on average). mRNA from each sample was reverse transcribed into cDNA and compared directly to a pool of unsynchronized *S. cerevisiae* (YPS183) cDNA in a common reference design.

cDNA from each sample was hybridized to 2 dye-swapped microarrays, which contain at least two replicate spots for each oligo, yielding at least 4 technical measurements of expression intensity for each gene for each strain at each timepoint. Similarly, cDNA corresponding to unsynchronized mRNA from each of 23 diploid mutation accumulation (MA) lines was collected. Each of these samples was hybridized to 2 dye-swapped microarrays of the same design, along with the unsynchronized reference cDNA. In total 377 time-series microarrays were produced for the 10 natural strains, and 50 microarrays were produced for the unsynchronized MA lines.

Data were quantified, filtered, and normalized, resulting in expression measurements for an average of 5879.9 genes per strain (92.4% of the genome) with an average mean standard error (sem) of 0.175 per gene per strain per time. We also estimated a biological replicate error (sem) of 0.554 between 2 microarrays after independent synchronization, release, and sampling of 2 cultures of the same strain at 63 min.

Subsequent analyses were performed after 2 additional modifications of these data. We first excluded a set of 91 transposable (Ty) element genes. Then to compare expression levels across strains at identical CDC-developmental states, we calibrated each natural strain's gene expression trajectories to a common cell-cycle length of 267 min. We estimated each strain's CDC period ω_i using a damped sine-wave regression of budding index (BI) time-series measurements and then obtained a calibrated CDC time-series by multiplying each

clock-sampled timepoint T_t by the ratio of the strain's CDC period with a standard of 267 min: $T_t^{CDC} = \omega_i T_t / 267$. We then rescaled each strain's expression data by fitting a cubic spline to each gene's expression trajectory and resampled at the calibrated CDC timepoints: $T_1^{CDC}, \ldots, T_{18}^{CDC}$.

Yeast strain information

Woodland strains used in this study are heterothallic haploid MATa derivatives of homothallic diploid *S. cerevisiae* and *S. paradoxus* isolates previously collected from state parks in Pennsylvania and New Jersey, USA (*31*). Laboratory strain YPS183 (*HO*Δ*:kanMX, leu2*Δ) is a heterothallic haploid MATa derivative of BY4741 (derived from S288c). Mating-type switching was prevented by disruption of the HO endonuclease locus (YDL227C) by homologous recombination with a Kanamycin resistance cassette. See Table 3.1 for more details. MA lines are diploid and were propagated asexually for 600 generations from a *leu2*Δ Y55 ancestor (provided by C. Zeyl (*16*)).

CDC synchronization of yeast cultures

Yeast cells were inoculated from frozen stock and cultured in SD minimal medium at 30°C (225 rpm). Cultures were diluted into fresh SD the next day and upon reaching a culture density of $OD_{600} \approx 0.25$, α-factor mating pheromone was added to a final concentration of 4 μM. Cultures were then incubated ≈75 min. until synchronized. The state of synchronization was determined by the appearance of $< 10\%$ shmoos and $< 10\%$ budding cells, visualized by light microscopy. Cultures were released from synchronization by removing the α-factor: 2x wash with 4°C S medium (SD without dextrose) and resuspension of cell pellets with fresh 18°C SD medium. Approximately 25 ml of each culture were distributed

into 18 flasks and incubated at 18°C (225 rpm).

The sampling time course consisted of 18 samples, taken approximately every 19 min. (real time), starting at 0 min. (time of release from arrest) and ending at 345 min. (see Figure 3.2 for specific timepoints). Upon sampling each culture was placed on dry ice, mixed with 20 ml of 100% EtOH stored at -20°C in a 50 ml Falcon tube, inverted, and placed immediately into a -80°C freezer. The first sample (0 min.) was taken after all flasks were returned to the incubator. Incubation of cultures at 18°C in SD medium more than doubles CDC length, allowing a more accurate comparison of measurements across strains by reducing the impact of temporal sampling variation.

Microarray processing

Total RNA was extracted from each frozen cell culture sample using Qiagen's RNeasy Kit, following manufacturer's instructions. cDNA was prepared from 15 μg of each RNA sample using SuperScript III reverse transcriptase (Invitrogen), after which each cDNA was purified using the Invitrogen Dye Purification Module. The Corning UltraGAPS glass slide platform was used for all microarray hybridizations. In preparation each glass slide was spotted twice with each of 6360 DNA targets using the Agilent yeast 70-mer library; these oligos target the $3'$ end of mRNA transcripts. Hybridizations followed a common reference design, using RNA extracted from unsynchronized culture samples of laboratory strain YPS183 at OD_{600} 1.1. Each slide was hybridized with equal amounts of synchronized and common reference cDNA samples, coupled to either the 555 or 647 AlexaFluor fluorophores (Invitrogen). 2 dye-swapped technical replicate slides were produced at each timepoint for each strain. Hybridized slides were incubated for 24–65 hours at 42°C. Slides were prepared for scanning by serial incubation in wash buffers and dried using both a vacuum and high-purity, filtered N_2 gas.

Image quantification

Slides were scanned at 10 μM resolution, 100% power, using an Agilent GenePix 4000B scanning confocal laser microscope. Gain was adjusted manually to optimize signal–noise ratios and to balance channel intensity distributions. GenePix v6.0 (Agilent) was used for feature identification and subsequent spot quantification per channel. Bad spots were flagged manually for removal from further analysis; otherwise all spots were treated equally in GenePix.

Within-slide data normalization

Each microarray's spot intensity data were normalized using custom Python software with calls to R. The median statistic was used to calculate foreground and background intensity values for each channel of each spot, yielding per-channel intensity distributions X^R and X^G and corresponding background distributions B^R and B^G. Each intensity distribution was log_2 (lg) transformed and scaled to remove multiplicative gain effects:

$$Y^R = \quad lg(X^R) - f(lg(B^R)) \tag{3.1}$$

$$Y^G = \quad lg(X^G) - f(lg(B^G)). \tag{3.2}$$

$f(lg(B^R))$ and $f(lg(B^G))$ are the global gain estimates for each channel, estimated as the 20^{th} percentile of the distribution over the local spot background intensities. To avoid erroneous quantification due to high local background intensity, a spot's local background was instead subtracted if this local background significantly deviated from the global background distribution ($>$ 3.5 SD from mean). A spot was discarded if both channel intensities

became non-positive after this correction.

The 2 intensity distributions over remaining spots were then mean centered:

$$Y_{\mu}^{R} = \quad Y^{R} - \mu(Y_{R}) \tag{3.3}$$

$$Y_{\mu}^{G} = \quad Y^{G} - \mu(Y_{G}). \tag{3.4}$$

To remove non-linear dependencies due to die-bias, each distribution was transformed using lowess regression (implemented in R) on each print-tip group, with a span of 0.3 and 2 iterations, using the following:

$$M = \quad Y_{\mu}^{R} - Y_{\mu}^{G} \tag{3.5}$$

$$A = \quad \frac{Y_{\mu}^{R} + Y_{\mu}^{G}}{2} \tag{3.6}$$

$$M' = \quad \text{lowess}(A, M, span, iterations) \tag{3.7}$$

$$Y_{L}^{R} = \quad Y_{\mu}^{R} - M'/2 \tag{3.8}$$

$$Y_{L}^{G} = \quad Y_{\mu}^{G} - M'/2. \tag{3.9}$$

The median of the lowess corrected log ratio distribution M' was then used to center each channel's intensity distribution:

$$Y_{L_c}^{R} = \quad Y_{L}^{R} - median(M')/2 \tag{3.10}$$

$$Y_{L_c}^{G} = \quad Y_{L}^{G} + median(M')/2. \tag{3.11}$$

A single, normalized distribution of log ratio spot values was computed by subtracting the reference channel from the synchronized channel, e.g. $M^{pre} = Y_{L_c}^R - Y_{L_c}^G$. Replicate spots within each microarray were averaged to generate a single intensity value for each unique oligo, comprising a distribution M of up to 6360 log ratio values per microarray slide (assuming no missing values).

Among-slide scale normalization

To compare intensity values across slides i, scale normalization was applied to each slide distribution M_i. Each slide distribution was modeled as a mixture of a Gaussian distribution (to scale the bulk of lowly expressed genes) and an empirical-null biological distribution (to scale the extreme intensity tails). This was implemented as a smooth transformation of each intensity distribution, using a quadratic polynomial, which was applied piecewise to the positive and negative halves of M_i:

$$Z_i^L = b_i M_i + b^L M_i^2 + \epsilon_i \tag{3.12}$$

$$Z_i^R = b_i M_i + b^R M_i^2 + \epsilon_i \tag{3.13}$$

$$Z_i = Z_i^L \cup Z_i^R. \tag{3.14}$$

To ensure that each transformed distribution had a smooth transition through 0, a single slide-specific slope b_i was used for both halves of the distribution. Thus three parameters b_i, b^L, b^R are needed to scale/transform a distribution. b_i was estimated for each slide i as the inverse slope from linear regression of the inner 80% of each M_i to a centered Gaussian distribution $N(0, \sigma)$, which served as a common reference distribution for all slides. This linear correction (rotation of a curve on a percentile-percentile plot) adjusts the

range for the majority of intensity values and consequently can inflate the range of extreme intensity values. To avoid this each tail was adjusted independently using a quadratic term, obtained for each half-distribution by quadratic regression constrained using the slope b_i (the negative half-distributions were rotated 180° clockwise before transformation).

To estimate the three universal parameters (σ, b^L, and b^R), we used a set of biological replicate hybridizations of YPS183 (described below). In principle these data should not contain differentially expressed genes, and so they constitute a suitable empirical-null (control) distribution. σ was estimated as the average σ_j of each control distribution before scale normalization. b^L and b^R were estimated as the average b_j^L and b_j^R obtained by regressing the control distributions against the reference Gaussian distribution. This transformation procedure was applied independently to each experimental distribution.

In addition gene expression values should generally correlate well over time, for a given strain. This was accounted for implicitly, by estimating the quadratic coefficients b_j^L and b_j^R using only the 1000 least temporally variable genes, which were selected using the synchronized expression data by ranking each gene by its expression variance over CDC timepoints for each strain, and taking the average of these variances.

Reference channel bias correction

To ensure that all microarrays were hybridized using freshly extracted RNA, we used 3 separate batches of common reference RNA (denoted by α, β, and γ), obtained from separate cultures of unsynchronized YPS183. Thus each batch serves as a biological replicate of same condition. cDNA from one of these 3 batches was hybridized onto each microarray along with a synchronized cDNA sample. To correct for potential bias on these synchronized microarrays due to biological replicate variance on the reference channel, one batch (α) was selected as the standard reference sample, and microarray data for β and γ were

calibrated against its distribution. To estimate the bias for each gene in the β and γ batches, a set of 16 microarrays were processed to compare the different references directly. 10 slides were processed using one pair of batches (β vs. α) and 6 slides for the other pair (γ vs. α) using dye-swapped technical replication. These data were normalized as described above.

Using these data, synchronized gene expression data were calibrated with respect to the α standard reference. A correction factor was estimated for each gene in both the β and γ batches as the average log ratio for that gene across the replicate slides of that batch, $< lg(X^R_\alpha / X^R_\beta) >$ and $< lg(X^R_\alpha / X^R_\gamma) >$. Thus each correction factor estimates the deviation in expression level between one batch and the α reference batch. These correction factors were then subtracted from the log ratio of each measurement M_{ik} for each strain i at each timepoint t (depending on whether its reference data came from the β or γ batch):

$$M^{cal}_{it} = lg(X^G_{it}/X^R_\beta) - < lg(X^R_\alpha/X^R_\beta) > \quad (3.15)$$

$$M^{cal}_{it} = lg(X^G_{it}/X^R_\gamma) - < lg(X^R_\alpha/X^R_\gamma) > . \quad (3.16)$$

Missing data imputation

In order to align gene expression trajectories across strains and to apply the singular value decomposition (SVD) for expression data analysis, it was necessary to estimate the missing intensity values (overall average of 7.6% values). We imputed values for each strain separately, after removing genes which were missing more than 50% of timepoints. Each missing value was replaced by the weighted average of a K-nearest neighbors estimate and an estimate based on cubic spline interpolation of the gene's expression trajectory (with 18 knots):

$$M_{it}^{cal} = \alpha \text{KNN}(k) + (1 - \alpha)\text{CS}(18) \tag{3.17}$$

We found that $k = 20$ and $\alpha = 0.9$ minimized the imputation error of an independent, time-series data set (culture of a yeast laboratory strain arrested with α-factor) (*32*).

In addition, insufficient RNA was extracted for 7 samples (there is no microarray data for these samples); this affects 3 strains (YPS2060, timepoints 111, 152, and 194; YPS2067, timepoints 87, 111; YPS3137, timepoints 63, 227). To recover data for these timepoints, we imputed each gene's missing values by fitting a cubic spline to each gene expression trajectory and recovering the interpolated values at each missing timepoint.

Morphological calibration of expression trajectories

While the temporal ordering of developmental states is consistent among strains, the rate of CDC progression is strain-specific, so that comparing expression levels sampled at identical clock timepoints may reflect spurious variation. To control for this constant strain-specific variation, CDC gene expression profiles were calibrated by aligning the budding index (BI) profiles of each strain using a simple linear scaling procedure. Here the BI serves as a proxy for developmental state.

Budding measurements were obtained from cell samples taken following the synchronization procedure described above. Sampling consisted of fixing cell culture in 3.7% formaldehyde and storing at 4°C. Each sample was then stained with DAPI and mounted on a glass microscope slide. The number of cells having two distinct nuclei was counted under DAPI-filtered fluorescence light, assuming that under normal light, a single identifiable bud was found attached to a mother cell. At least 200 cells were counted on each slide.

Define $BI(t)$ as the proportion of budding cells divided by the total number of cells counted. To estimate CDC length ω_i, a 5-parameter damped sine regression was performed on the BI profile for each strain i (*33*):

$$BI_i(t) = y_{0i} + A_i e^{-t/D_i} sin(\frac{2\pi t}{\omega_i} + C_i) \tag{3.18}$$

To avoid extrapolation issues, each BI profile was aligned to that of the laboratory strain YPS183, which displayed the longest CDC period at 267 min. A single scale parameter was estimated for each strain (excluding YPS183) as the ratio of the standard strain's period (267) to the target strain's period. A cubic spline was used to approximate each profile as a continuous curve, constraining the curve to pass through knots defined by the BI values at each timepoint. A strain's BI profile was then re-evaluated at the CDC calibrated timepoints: $T_t^{CDC_i} = \omega_i T_t/267, \forall\ t \in 1 \ldots 18$. Since all ratios were ≤ 1, this resulted in retarding natural strain profiles relative to the laboratory strain (Figure 3.36).

Finally we rescaled each strain's expression data by fitting a cubic spline to each gene's expression trajectory and resampling at the calibrated CDC timepoints for that strain, T^{CDC_i}.

Estimation of CDC-phases

CDC-phases were determined using data from (*34*), where the fraction of budded cells (F_{S+G_2+M}) and the fraction of post-S-phase cells (F_{G_S+M}) were estimated using the S288c strain grown in minimal medium at 18°C. 267 min. was used as the CDC length based on microscopy of YPS183 (see above). From this, the G_1/S transition occurs at 47 min., G_2/M occurs at 97 min., and M/G_1 occurs at 267 min. The remaining transition, G_2/M, was placed at 218 min using our gene expression data for the lab strain YPS183 (Figure 3.34). At timepoint 218 min., expression of the Y19-kinase Swe1 (YJL187C, inhibitor of Clb2-

Cdc28 activity) peaks in G_2 (*35, 36*). This timepoint is coincident with nearly maximal levels of the B-type cyclin Clb2 (YPR119W) and low expression of the Y19-phosphatase Mih1 (YMR036C), which promote G_2/M progression (*37, 38*). Cdc2 (YDL102W), a gene whose deletion arrests CDC progression at the G_2/M checkpoint, also shows nearly maximal expression at this time.

Estimation of evolutionary variation in gene expression

Natural variance in gene expression was estimated for 6,263 genes i at 18 timepoints t using a first-order Taylor series expansion of the variance function to remove sampling noise. Dividing each variance value by n_{gen}, the estimated number of generations until coalescence of these strains (*39*) yields the per-generation increase in expression variance $\sigma_n^2(i,t)$, a statistic which reflects the composition of the evolutionary forces of mutation, drift, and selection. To calibrate the natural variation, $\sigma_n^2(i,t)$ was scaled by the expected gene expression variance per-generation under mutation–drift equilibrium $\sigma_m^2(i)$ using MA line expression measurements. This yields an F-ratio of natural genetic variance to neutral genetic variance, per-generation, scaled by respective degrees of freedom:

$$F(i,t) = \frac{\sigma_n^2(i,t)}{\sigma_m^2(i)} \times \frac{22}{8} \tag{3.19}$$

Model for natural expression variance

Natural variance in gene expression was estimated as follows. In order to correct for temporal culture sampling noise which could inflate the true variance at each sampled time t, a correction was applied to the sample variance over natural strains $\sigma_n^2(i,t)$. Define the normalized expression level of gene i from strain j at sampled time t as $Y_{ij}(t)$. This can be modeled as an unknown time-indexed function $Y_{ij}(t) = f(t)$. The first-order Taylor series

expansion of this function is

$$f(t) \approx g(\tau) + g'(\tau)(t - \tau) + \delta\tau, \tag{3.20}$$

where $g(\tau)$ is the true (unknown) expression level at true time τ, $g'(\tau)$ is its derivative, multiplied over the time interval $(t - \tau)$, and $\delta\tau$ is the residual. Variance in gene expression sampled at the same time across strains can be estimated as the variance of this approximation, using the Delta Method (*40*):

$$Var\{f(t)\} \approx g'(\tau)^2 \times Var\{t\} + Var\{\delta\tau\}, \tag{3.21}$$

where $Var\{\delta\tau\}$ is the desired estimate of natural expression variance and $Var\{t\}$ is an unknown constant representing the temporal sampling variance of different cultures (sampling noise or jitter). We solved for $Var\{\delta\tau\}$ as the difference between empirical expression variance across strains and the sampling variance:

$$Var\{\delta\tau\} = Var\{f(t)\} - g'(\tau)^2 \times Var\{t\}. \tag{3.22}$$

We reasoned that $Var\{t\} = 267/64.0 = 4.17$, that is $1/8^2$ of the calibrated CDC length, where $1/8$ corresponds to a maximum CDC-phase deviation between strains sampled around true time τ ($267/8 = 33.9$ min.). Thus the per-generation natural expression variance $\sigma_n^2(i,t) = Var\{\delta\tau\}/n_{\text{gen}}$.

Discrete derivative estimates of every gene expression measurement were computed as

$$g'(t) = (Y_{ij}(t) - Y_{ij}(t - \Delta t))(t - \Delta t)^{-1}, \tag{3.23}$$

where Δt constitutes a time delay between successive measurements. A single Δt was estimated from cubic spline interpolated profiles Y_{ij} of each gene for each strain by maximizing the correlation between the expression profile and its derivative profile, evaluated over $\Delta t \in \{10, ..., 140\}$ min. This range of possible values was chosen for biological propriety: we anticipated recovering noisy derivatives under 10 min. and irrelevant derivatives much beyond half of the length of the aligned cell cycles (134 min.).

Estimation of the number of generations until coalescence

The number of generations until coalescence of natural yeast strains, n_{gen}, was estimated as 8,342,391 generations, using $\frac{E\{Substitutions/site\}}{2\mu}$, where $\mu = 1.84 \times 10^{-10}$ is the mutation rate per nucleotide site per generation (*41*) and $E\{Substitutions/site\} = 0.00307$ was determined by maximum likelihood phylogeny estimation with a molecular clock (DNAMLK) (*42*), using 36 intron sequences from 2 natural woodland strains and the laboratory strain of *S. cerevisiae*: S288c, YPS128, and YPS606. Sequences of the 2 natural strains were obtained from cells isolated in the same locations as those whose transcriptomes were measured. We obtained the assembled genome sequences from (*43*) (http://www.sanger.ac.uk/Teams/Team118/sgrp/) and identified intronic regions using custom gene mapping software and exon annotations for the laboratory strain genome (*44*) (http://www.yeastgenome.org).

To account for error in the estimate of coalescence time, we calculated a 95% confidence interval around n_{gen} by treating the number of nucleotide substitutions as a Poisson-distributed random variable with parameter $\lambda = 0.00307 \times 50,185 = 154.06795$ sites, the estimated total number of intronic substitutions. This interval is 7,010,869 to 9,728,260 generations.

Mutational variance estimation

F-ratios were only computed for the subset of genes having significant mutational variance among MA lines. These genes were determined by a linear mixed model analysis of each gene i using Python with calls to PROC MIXED in SAS software (*45*):

$$Y_{ijkm} = \mu_i + Line_j + Spot_m(Slide_k) + \epsilon_{jkm}, \tag{3.24}$$

where Y_{ijkm} is the log_2 measurement of a given gene i from MA line j on slide k for spot m, μ is the grand mean, $Line_j$ are random line effects, $Spot_m(Slide_k)$ are random slide-specific spot effects, and ϵ_{jkm} are line-specific residuals. These effects are normally distributed with a mean of 0 and variance estimated as follows:

$$Line \sim N(0, \sigma^2_L); Spot(Slide) \sim N(0, \sigma^2_{S(k)}); \epsilon_{(j)} \sim N(0, \sigma^2_{\epsilon(j)}) \tag{3.25}$$

Mutational variance $\sigma^2_m(i)$ was estimated as $\sigma^2_L/2n_{\text{gen}}$, where $n_{\text{gen}} = 600$ is the number of generations over which the MA lines were propagated. The factor of 2 is used here to account for asexual diploid evolution (*39*). Using a false-discovery rate (FDR) multiple test correction of 0.25, we identified 4973 genes with significant mutational variance estimates.

Gene enrichment analyses

Enrichment analyses of genes grouped by Gene Ontology (GO) terms or life-cycle related terms were performed using Fisher's Exact test and evaluated using a FDR of 0.05 on the number of groups/terms, unless otherwise specified. Background frequencies were determined from the union of all genes considered in a given analysis.

The set of GO terms consisted of 88 terms (GO-Slim process, function, component)

(www.geneontology.org, accessed 29 Dec, 2007). The set of 8 life-cycle related terms was compiled from various sources. In particular, periodic genes refer to a set of 822 genes identified from 5 yeast CDC time-series data sets (*46*) as being generically periodically expressed with the CDC. CDC refers to the subset of these periodic genes which are also labeled by the GO-Slim term 'cell cycle'. Transcription regulator refers to the genes labeled by the GO-Slim term 'transcription regulator activity'. Remaining terms refer to GO-Slim terms of similar name.

Multivariate transcriptome analyses

Various analyses were performed using custom Python software with calls to R, SAS (*45*), and Mathematica (*47*).

Global and time-specific eigenvectors

Singular value decomposition (SVD; implemented in R) was applied to both the entire *S. cerevisiae* CDC data set Y (after mean centering), consisting of the common set of 6082 genes with complete data measured in 9 strains across 18 timepoints (6082×162 matrix). SVD was also applied to subdivisions of this data set, the 18 independent expression matrices $Y(t)$ with 9 conditions each (the 9 strains):

$$U, \Sigma, V^t = \text{SVD}(Y). \tag{3.26}$$

Eigenvectors $u_i \in U$ comprise the orthonormal basis of the entire data set Y and are referred to as global eigenvectors, while those corresponding to covariation at particular timepoints are referred to as time-specific eigenvectors $U^r(t)$ or CDC-directions of covari-

ation having rank r (corresponding to eigenvalue magnitude).

In order to compare CDC-directions properly, they must derive from the same variable space, so the 18 expression matrices $Y(t)$ were first projected onto the 162-dimensional global eigenvector matrix U to recover a set of 18 eigengene expression matrices (EEMs) as follows:

$$Y'(t) = U^t \cdot Y(t). \tag{3.27}$$

SVD was then performed on each $Y'(t)$ to recover a time-specific eigenvector matrix $U(t)'$ containing 9 column vectors. We denote the largest such vector as the major or preferred CDC-direction and the 8 remaining, smaller vectors as minor CDC-directions.

Random angle distribution for testing distance between CDC-directions

To assess the significance of angles between CDC-directions, a null distribution of angles was generated by comparing all pairs of CDC-directions $\angle U^r(t_i)\ U^s(t_j)$, resulting in $\binom{18}{2}$ time combinations $\times\ 9^2$ rank combinations $= 12,393$ values. Composition of this distribution thus randomizes the choice of rank and time. In all calculations the reported angles are reflected around 90°, since eigenvectors are undirected.

CDC-direction projection plot

We constructed a 2-dimensional projection plot to summarize the change in the angle between major CDC-directions (in 162 dimensions) for each timepoint. First, we computed the angles between all CDC-direction pairs. Then, starting from the first timepoint (0 min., shown at 12 o'clock in 2 dimensions; Figure 3.12A), we computed the placement of the

next vector by minimizing the sum of squares error between the 162-dimensional angle and the 2-dimensional angle formed between the current eigenvector i and the 3 eigenvectors at preceding timepoints $i-1$, $i-2$, and $i-3$:

$$Error_i = (\Theta_{i,i-1} - \theta_{i,i-1})^2 + (\Theta_{i,i-2} - \theta_{i,i-2})^2 + (\Theta_{i,i-3} - \theta_{i,i-3})^2 \tag{3.28}$$

Numerical optimization was implemented by the simulated annealing algorithm in Mathematica.

F-ratios of multivariate evolutionary covariation along CDC-directions

To recompute F-ratios representing multivariate evolutionary covariation along CDC-directions, we projected gene expression data onto the (1 major and 8 minor) CDC-directions at each timepoint and computed the resulting sample variance over strains. This was performed for both natural and neutral expression data. The first-order Taylor series approximation (see above) was applied to the natural variance values. Since neutral data were taken from batch culture, we took the average of neutral variances over time for each degree of eigenvector (major, 2^{nd}, ..., 9^{th}). F-ratios $F^r(t)$ were computed as the ratio of natural variance at each timepoint t to the (time-averaged) neutral variance, for each rank r, scaled by degrees of freedom (8 and 22, respectively).

Linear mixed models factor analysis

To determine the minimum number of independent latent (underlying) factors consistent with observed covariance matrices, EEMs were analyzed using a linear mixed model with mean and a strain-specific random eigengene effects taking a factor analytic variance-

covariance structure G, composed of q factors (specified using "type=fa0(q)" in `PROC_MIXED` in SAS). For EEMs corresponding to natural strains, q took on values from 0 to 9. For the MA line EEMs q ranged from 0 to 23. Models were evaluated by a step-up procedure, where likelihood ratio tests were computed between successive increases in q. The number of factors reported is $q + 1$, where q corresponds to the first model yielding a significant likelihood ratio test, going in decreasing order from $max(q)$ to $min(q)$.

Canonical correlation analysis

Canonical correlation analysis (CCA) was used to compare gene expression matrices between CDC timepoints within *S. cerevisiae* (using the CCA package in R). In contrast to SVD analysis, where data at each timepoint were decomposed independently for subsequent comparison of eigenvectors, CCA was used to compare data between 2 timepoints directly to identify vector directions of maximal similarity. Significance of the primary canonical variate was determined using Wilks' Lambda distribution.

For each of $\binom{18}{2} = 153$ unique pairs of timepoints s and t, CCA was applied to the 2 corresponding EEMs X_s and X_t) of size 162×9. Computing canonical correlations requires inverting each EEM (as $X_s'X_s$ and $X_t'X_t$), so we needed to apply a 2 parameter regularization prior to CCA, by estimating λ_1 and λ_2 such that $X_s'X_s+\lambda_1 I_s$ and $X_t'X_t+\lambda_2 I_t$ are invertible (I_s and I_t are identity matrices). Regularization was implemented by the CCA package in R.

Common principle components test

The common principle components (CPC) test (*19*) evaluates a set of nested hypothesis tests (the Flury hierarchy) using a likelihood ratio Chi-squared statistic to determine the

number of eigenvectors shared among 2 or more matrices (ranging from 0 up to the number of columns in the matrices - 2); whether or not the eigenvector lengths are proportional to each other (Proportional or CPC); or whether all eigenvectors and their lengths are equivalent, indicating identical eigenstructure (Equivalent). Natural covariance matrices were compared using the CPC test between sequential timepoints. All CDC-expression data (10 strains, 18 timepoints) were first projected onto top 9 global *S. cerevisiae* eigenvectors and then split by timepoint to recover 18 9×9 eigengene expression matrices (EEMs). Note that there are only 9 degrees of freedom for the 10 columns of data at each timepoint. 9×9 covariance matrices were then computed from these EEMs. These covariance matrices represent less than 50% of the total variation at each timepoint. To evaluate the results of the CPC test, the Chi-squared test statistic was used, which reflects the log likelihood ratio comparing one of the nested alternative models against the null model of unrelated eigenstructure (jump-up approach). The model selected for each comparison was the smallest one yielding a significant p-value ($P < 0.05$).

Comparison of genome-wide CDC coexpression patterns

Genome-wide coregulatory structure evolution

In order to compare the broad, genome-wide pattern of temporal coregulatory structure across yeast strains, gene correlation matrices were computed for each strain and compared between strains. 10 6082×6082 correlation matrices were computed using Spearman's rank correlation coefficient, based on the 10 6082×18 CDC-expression matrices. Mantel's R matrix correlation coefficient was calculated between each pair of correlation matrices, yielding a 10×10 similarity matrix relating the strains. P-values were estimated for each comparison using the Mantel test with 100 matrix permutations, and overall significance

was determined at FDR < 0.01. To confirm that the genome-wide results were not due to an averaging effect, a similarity matrix was computed using a subset of 270 transcription regulator genes (corresponding to genes labeled by the GO-Slim term 'transcription regulator activity'). To provide a negative control, a null similarity matrix was also computed from correlation matrices produced from row- and column-permuted expression matrices of each strain.

Positive controls were generated by comparing data generated through a parametric resampling procedure to our CDC-expression data for strain YPS183. One of the largest sources of error in comparing time-series data between strains occurs in the sampling of RNA at each prespecified timepoint $\tau_i, i \in \{1, \ldots, 18\}$, where the actual RNA sample is taken at some timepoint t_i around τ_i (see 'Estimation of evolutionary variation in gene expression', above). Using our estimate of 4.17 as the sampling variance around τ_i, we synthesized CDC-expression data for YPS183 by randomly resampling the "observed" timepoints t_i using a Normal distribution with mean τ_i and variance 4.17. We fit each gene's expression trajectory to a cubic spline function with 18 knots at the prespecified timepoints τ_i and then reevaluated each trajectory at the 18 resampled timepoints t_i. To control for error due to biological replication, we also generated 10 matrices where, in addition to time sampling noise, Gaussian measurement noise was added to every data point (expression level of gene g at timepoint t) using mean y_{gt} and variance 0.307 (corresponding to the square of our estimated biological replicate error 0.554).

Temporal coregulatory structure evolution

To identify whether temporal coregulatory structure evolves differentially across the CDC, each strain's expression data set was partitioned into 3 overlapping subsets of 9 timepoints (1–9, 5–13, and 10–18, denoting early, middle, and late CDC-phases). This generated

$10 \times 3 = 30$ data sets. Gene correlation matrices were computed as described above, and a 30×30 similarity matrix was computed, reporting Mantel matrix correlation coefficients. Hierarchical clustering of this similarity matrix (overall and for each phase subset) was performed using average linkage and Pearson correlation.

Comparison of phase-directions of covariation

SVD was performed on each mean-centered phase subset above, and the major eigenvectors (phase-directions) were extracted. Angles corresponding to distances between these major phase-directions were computed in 2 ways. First the early, middle, and late phase-directions were compared within each strain (among phases within strain), to identify variability in the major temporal direction of variation across the CDC. Next the early, middle, and late phase-directions were compared to the corresponding phase-directions of other strains (within phase among strains), to identify evolutionary variability in the major temporal direction of variation at each CDC-phase. Each angle was compared to a random angle distribution (see above) and significantly small angles identified (FWER < 0.05).

Analysis of modular coregulatory structure evolution

Comparison of whole-genome correlation matrices represents a high-level view of regulatory structure evolution. To obtain a local, modular view of each strain's regulatory structure, (overlapping) gene clusters were computed for each gene in each strain, using Pearson correlation to compare the CDC-expression profiles between genes. Let X_{is} be the CDC-expression data for the i^{th} gene in the s^{th} strain. The correlations of X_{is} with all other genes in the s^{th} strain were first computed. The top k most correlated genes were then designated as the k-module $M_{is}^{(k)}$ for gene i in strain s. This was repeated for all genes

in all strains, yielding $M^{(k)}$, a k-modular representation of transcriptome organization for all strains. Intersections between $M_{is}^{(k)}$ and $M_{it}^{(k)}$ were determined for each gene i for all pairs of strains s and t. The size of each intersection was divided by k to yield the module overlap proportion for each gene. Significance was assessed for each k-module $M_{is}^{(k)}$ by comparing it to a null distribution of overlap proportions, composed of 250 randomizations of the CDC-expression data for that gene and strain X_{is}. A p-value cutoff of 1/250 (0.004) was used for each k-module.

To assess the excess in the observed amount of module overlap between strains, we considered the expected overlap based on binomial sampling. Given a set of k genes for one strain (from a genome of n genes), the probability that 1 gene from a second strain intersects with this set of k is k/n. The expected overlap, given the number of trials k and the success probability p, can then be modeled as $kp = k \cdot k/n = k^2/n$. After scaling this statistic by k to represent the expected proportional overlap, we recover the probability k/n. Thus the expected overlap proportion increases linearly over the range of module sizes by proportion of the genome, that is over $[0, 1]$ (see Figure 3.18).

Time-domain transformation model

A regression model was developed to test explicitly whether changes in a gene's CDC-expression between strains can be explained by a time-domain transformation. This model was applied to each gene i between all pairs of strains u and v. In this model (H_1) 3 time-domain parameters (α, β, and γ) are incorporated to a standard linear regression model (H_0):

$$H_0 : \quad X_{iv}(t) = A + BX_{iu}(t) + \epsilon \tag{3.29}$$

$$H_1 : \quad X_{iv}(t) = A + BX_{iu}((Beta(t, \alpha, \beta) + \gamma) \bmod 1) + \epsilon. \tag{3.30}$$

Given 2 shape parameters, the Beta cumulative distribution function generates a smooth, continuous, and invertible transformation curve between 2 time domains. Timepoints t are defined from 0 to 1 and represent the fraction of CDC progression. The parameter γ allows for a phase offset between time domains. Each mapped timepoint must be modulated around 1, so that the time-domain transformation is defined with respect to a single cell-division cycle. A linear regression model, depicting a time-independent fit of expression data, was used as the null model. This null model is a special case of the time-domain model, using $\alpha = 1$, $\beta = 1$, $\gamma = 0$, where $t = Beta(t, 1, 1) + 0$.

Since the time-domain model is defined with respect to a single cell-division cycle while our empirical data spans 1.3 cycles, the first 14 of the 18 timepoints were used for the expression trajectories each gene in every strain; the last 4 timepoints of data were averaged with the first 4 timepoints before model fitting. Sampled timepoint 14 (260 min., closest sampled timepoint to the calibrated CDC period of 267 min.) was subsequently used as the CDC period. Next a periodic cubic spline was fit to each 14-timepoint gene trajectory and then reevaluated at 18 timepoints covering exactly 1 CDC. Next these expression data $X_{\cdot u}$ were Z-standardized (mean centered and scaled by standard deviation). This has the effect of emphasizing the fit between temporal expression patterns (regression slope becomes equivalent to Pearson correlation) and eliminates y-intercept estimation.

The heterochrony model parameters were estimated in two stages. First the 3 time-domain parameters were estimated by exhaustive enumeration, evaluating fit by the sum of squares error between query and target CDC-expression trajectories:

$$(\rho, \sigma, \psi) = \text{argmin}_{P,\Sigma,\Psi} \sum_t \left(X_{iu}((Beta(t, \alpha, \beta) + \gamma) \bmod 1) - X_{iv}(t)\right)^2. \qquad (3.31)$$

The exhaustive search was performed using a predefined grid of parameter values. Esti-

mates of α and β were bounded within $[1/3, 3]$ at a resolution of 0.1, and γ estimates were bounded within $[-260/2, 260/2]$ at a resolution of 5 min. Each search thus consisted of evaluating and ranking 40,768 parameter triples.

Subsequently linear regression was used to estimate B, given the time-transformed X_{iu} data and $A = 0$. Null and alternative models were estimated identically, except that parameter values for the former were fixed at ($\alpha = 1, \beta = 1, \gamma = 0$), while those of the null model use these only as initial values. Model significance was assessed using a likelihood-ratio F-test comparing the increase in variance explained by the time-domain model to its unexplained variance:

$$F = \frac{(R^2_{H_1} - R^2_{H_0})/(4-1)}{(1 - R^2_{H_1})/(18-4)} \tag{3.32}$$

P-values were computed using 3 and 14 degrees of freedom, respectively. Between each pair of strains, F-values were computed for all genes i, and significant genes were identified after multiple test correction (FDR < 0.05). This model is invertible when reversing the dependent and independent variables X_{iu} and X_{iv}, given the inverted parameter values $1/\alpha$, $1/\beta$, and $-\gamma$. Thus an optimal time-domain parameter triple was estimated for every gene over all $10(10-1)/2 = 45$ unique pairs of strains.

95%-equivalence timing curve ensembles

Although time-domain parameter estimation returns unique optimal values for each search, the top k_i^j parameter triples, representing an ensemble of equivalent time-domain transformation curves (timing curves), were retained for every gene i in each strain comparison j. Each timing curve ensemble of k_i^j triples was defined as the number of triples yielding the sum of squared error (SSE) error within 95% of the optimal model SSE. A single, global 95% equivalence bound was estimated through a parametric resampling procedure,

as follows. Expression trajectories for 10 genes were selected randomly. For each trajectory, 100 resampled trajectories were then generated, adding both expression level and time-sampling noise (see 'Genome-wide coregulatory structure evolution' above). Each resampled trajectory was then fit to its original trajectory using the heterochrony regression model, and the optimal SSE was obtained. The pool of 100 SSE values forms a null distribution over the time-domain transformation curves matching an expression trajectory to noisy resamplings of itself. The 95^{th} percentile of each gene's mean-centered SSE distribution was computed as its equivalence bound. Finally these 10 per-gene equivalence bounds were averaged, yielding 2.27, the value of the global equivalence bound. Thus each 95%-bounded equivalence set contains the k_i^j parameter triples whose SSEs are within 2.27 units of the optimal SSE.

Calculating distance between timing curves

The distance between the timing curves for 2 genes was defined as the minimum root mean squared error (RMSE) over all pairs of timing curves in each gene's 95%-equivalence timing curve ensemble.

Gap statistic analysis

Gap statistic analysis was performed on the summary timing pattern distance matrix. k-means clustering was used to obtain clusters of the 4998 genes using $2 \geq k \geq 26$ and 100 random restarts for each k. A gap statistic was then computed at each k, given the data matrix and the corresponding clustering. Support for a particular k was identified as the smallest k such that $gap(k) \geq gap(k+1) - 2sem(k+1)$ (*48*).

Clustering of timing pattern distance matrices

k-means clustering with $k = 7$ was used to group 4998 heterochronic genes into 7 timing pattern groups for each of the 45 strain comparisons. k-means was applied to a $4998 \times n$ coordinate matrix recovered by metric multidimensional scaling of each of the 45 4998×4998 timing curve distance matrices, where n is the number of coordinate dimensions required to recover 100% of the covariation in each distance matrix. Each k-means procedure was run with 100 random restarts.

Heterochronic gene interactions

Pairs of genes which co-cluster in more of the 45 strain comparisons than expected were identified using each gene's profile of timing pattern cluster labels. The total number of strain comparisons where cluster labels matched was counted for all pairs of genes. Significance of each pair's number of matching labels was evaluated as a binomial probability of the observed number of matching labels with 45 trials and a cutoff of $P < 10^{-4}$ per pair. Since gene clustering was performed independently in each of the 45 comparisons, the matching success probability was computed for each comparison as the frequency of each respective cluster label in that comparison. Thus binomial probabilities were calculated as the binomial coefficient times the product of success probabilities for the k matching comparisons times the product of failure probabilities for the $45 - k$ non-matching comparisons:

$$P\{\geq k/45 \text{ successes}\} \approx \binom{45}{k} \prod_{i}^{45} \pi_i^{z_i} (1 - \pi_i)^{\neg z_i} \tag{3.33}$$

Here $z_i \in \{0, 1\}$ indicate whether comparison i had a match. Assuming uniform cluster frequencies, the expected number of matches would be $45 \times 1/7 = 6.42$.

ANOVA on timing patterns

To test whether timing modules show different timing patterns, the optimal timing curves for the set of within-module genes associated with each module were pooled from all 45 strain comparisons. This set constitutes the data pool for each module. An ANOVA test was performed between the 7 timing modules.

Test of modular timeline variability across genes

To assess variability among the gene expression timelines in each timing module, a distribution of random variances was generated from groups of random timing patterns of the same size as each timing module. Each random timing pattern was generated by randomly selecting an α, β, and γ value from the empirical distributions of these parameter estimates genome-wide (Figure 3.21). The random distribution is comprised of variances from 250 random groups of timing patterns. Since a timing pattern is defined by the 3 parameters α, β, and γ, a statistic is desired that reflects the variability of these parameters as a whole. So we computed variance at each of 100 timepoints across the CDC and then took the average of these 100 variances (for both observed and expected variances), obtaining a single variance estimate across each group of timing patterns.

N.b., perhaps a more conservative test would be to generate a set of random timing curves and then cluster them using k-means (with $k = 7$); however our timing modules were not simply the result of a single clustering, but are a subset of genes that show significantly similar timing pattern changes across 45 strain comparisons and map to the same clusters. As repeating this procedure in the same manner for random timing curves is not expected to return any clusters, we resorted to evaluating modular timeline variability in the manner described, using a stringent false discovery rate cutoff of 0.001.

Test for evolutionary timeline variability of a gene

Each gene is associated with a set of optimal timing curves across 45 strain comparisons, 36 of which are between strains of *S. cerevisiae*, while 9 are between species. Variation in timing curves across comparisons for the same gene characterizes the degree of evolutionary change in expression timing for that gene across the CDC. Natural evolutionary timing pattern variation for a gene was computed as the average of 9 variances, each of which is the timing pattern variance over the 8 comparisons starting from one particular strain. For example, starting with YPS183, we have a timing pattern curve relating YPS183 to YPS2055, YPS183 to YPS2060, YPS183 to YPS2066, ..., and YPS183 to YPS3137. The 9 variances starting from each of the 9 *S. cerevisiae* strains are then averaged. We tested for significance of this average evolutionary variance by comparing it to a random variance distribution. The random variance distribution consists of 1000 variance values obtained in the same way as the natural variances, except that random timing curves were used in place of the true timing curves. Random timing curves were generated from the empirical distribution of optimal α, β, and γ parameters (see Figure 3.21).

Acknowledgements

We wish to acknowledge Helen Murphy, Cara Winter, Fan Ge, and Irmina Gawlas for assistance in processing microarrays, DNA sequencing, and measuring CDC length.

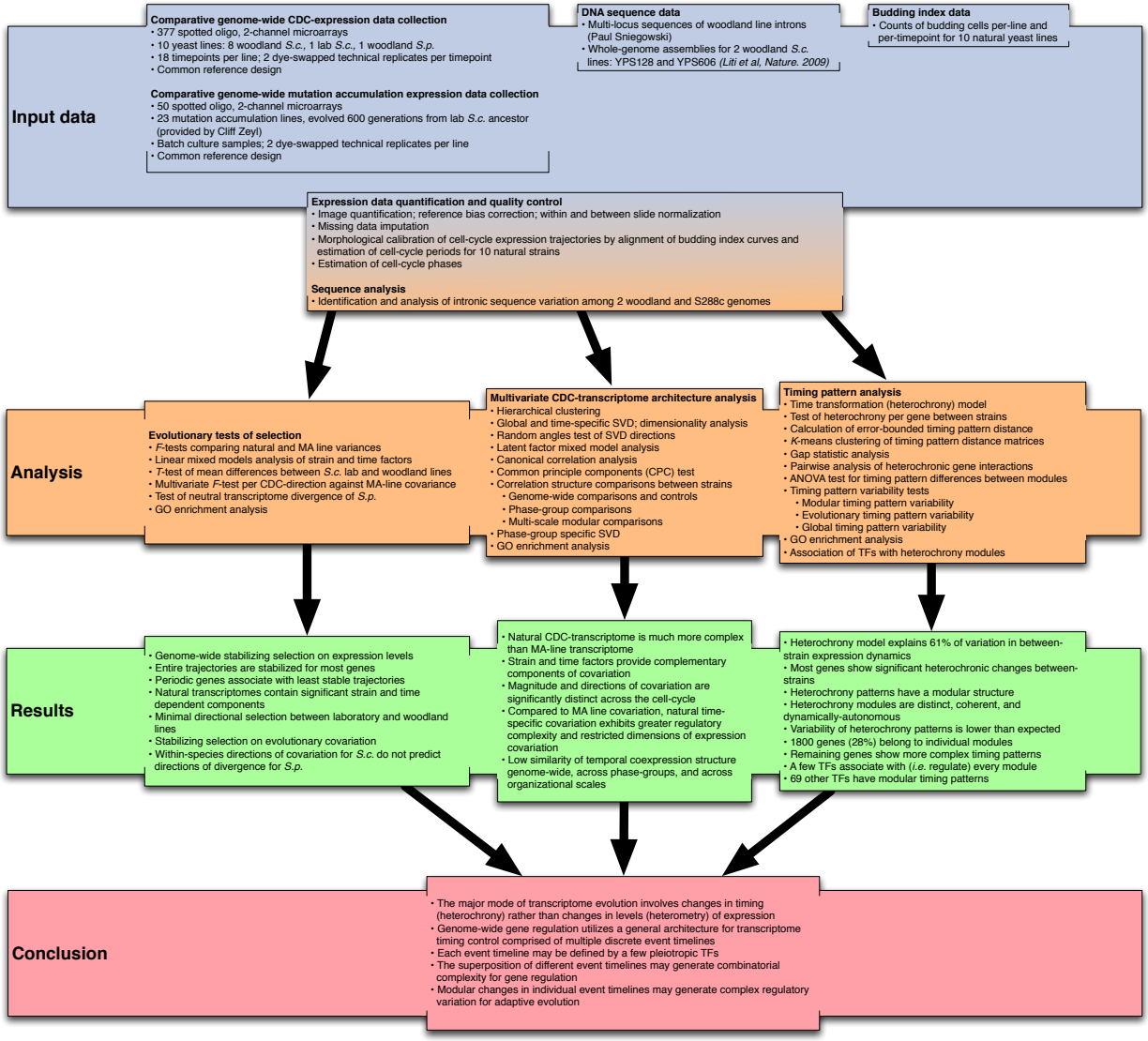


Figure 3.1: Overview of experimental design, analysis, results, and conclusion.

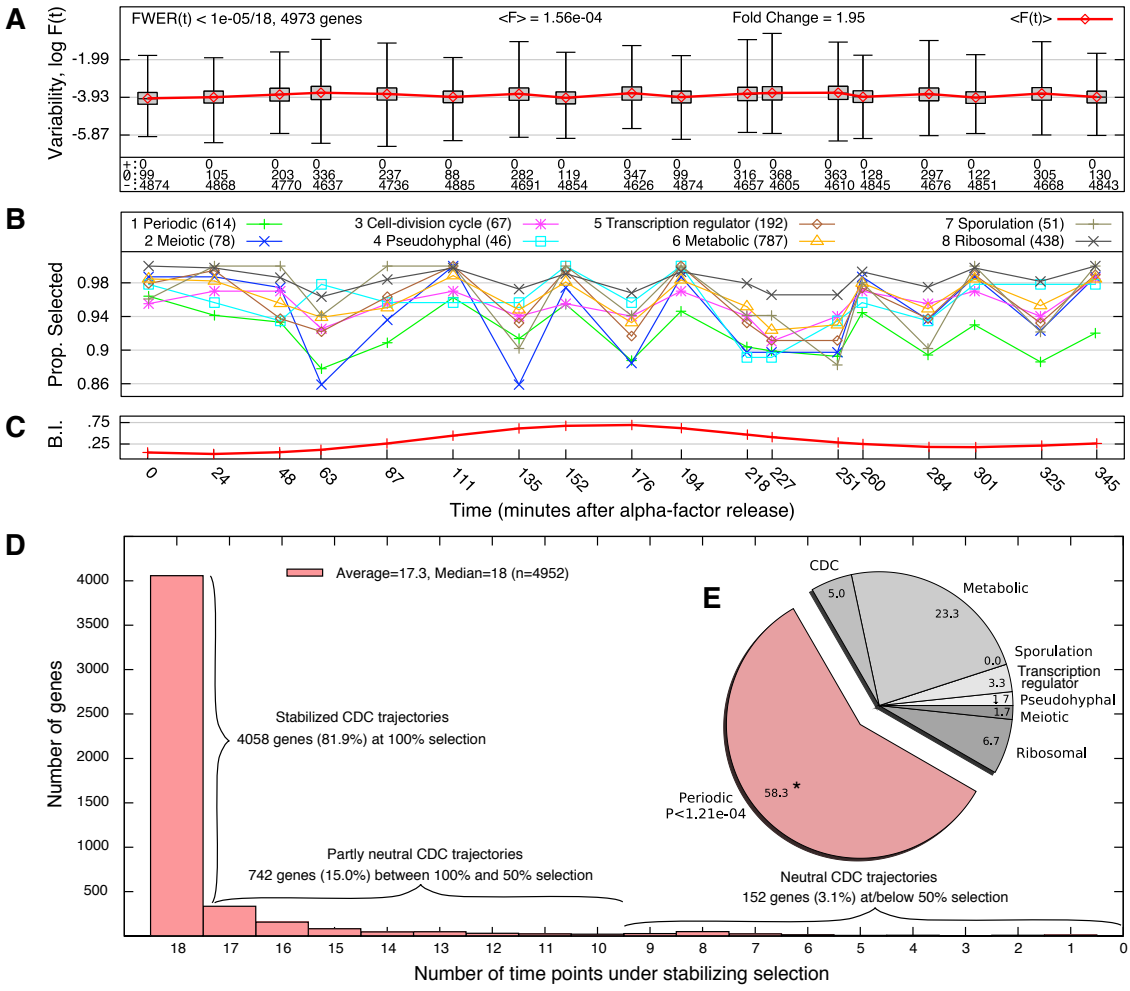


Figure 3.2: Genome-wide evolutionary gene expression variability among *S. cerevisiae* strains through the CDC

(A) Genome-wide evolutionary gene expression variability $F(t)$ among *S. cerevisiae* strains through the CDC, and the number of genes exhibiting positive (+), stabilizing (-), or no selection (0). Average variability profile (red), exhibiting a maximum fold change of 1.95. **(B)** Proportion of genes under stabilizing selection over time for 8 life-cycle terms, ranked by average proportion. The number of genes is in parentheses. **(C)** Average *S. cerevisiae* budding index. **(D)** Histogram of the number of timepoints for which a gene's CDC-expression trajectory undergoes stabilizing selection, separated into stabilized, partly neutral, and neutral categories. **(E)** Enrichment of life-cycle terms among neutral genes. *indicates significant enrichment (FDR < 0.05).

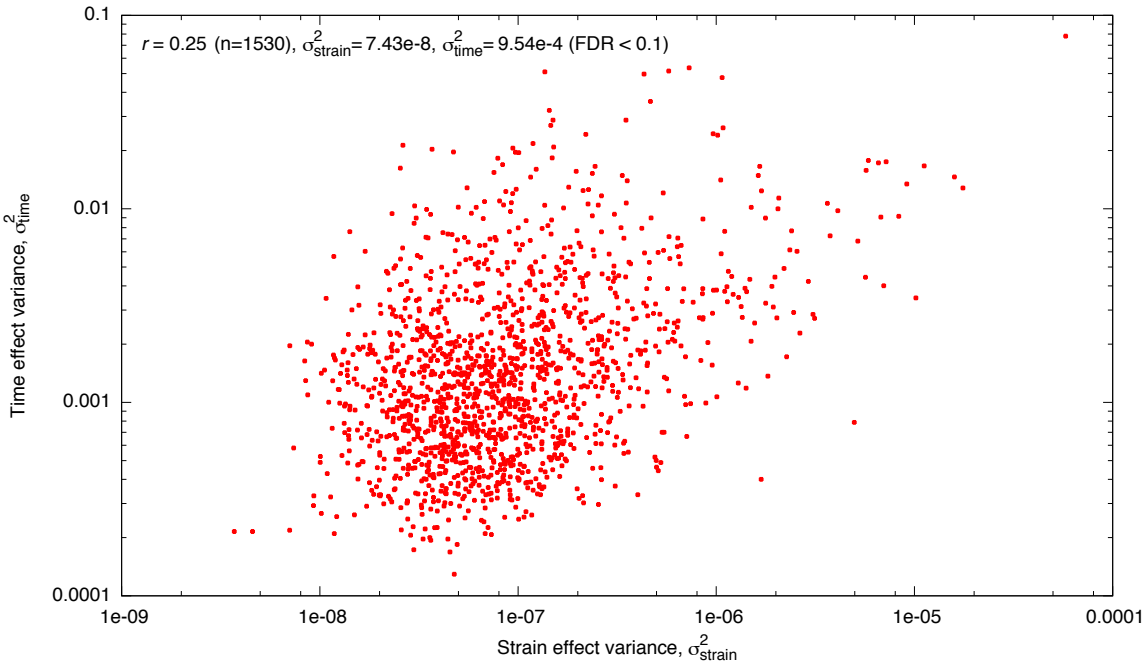


Figure 3.3: Relationship of two components of gene expression variation, time (temporal variation) and strain (strain divergence)

The correlation was computed among 1643 genes with significant time and strain effects in *S. cerevisiae* (FDR < 0.1 over all 6251×2 hypotheses). Strain variance components were corrected for sampling noise and mutational variance. Both time and strain variance components were scaled by degrees of freedom (17 and 8, respectively). Summary statistics **(top)** were calculated as the median for each effect using estimates of significant genes. 113 of the 1643 genes were omitted due to insignificant estimates of mutational variance at FDR < 0.25.

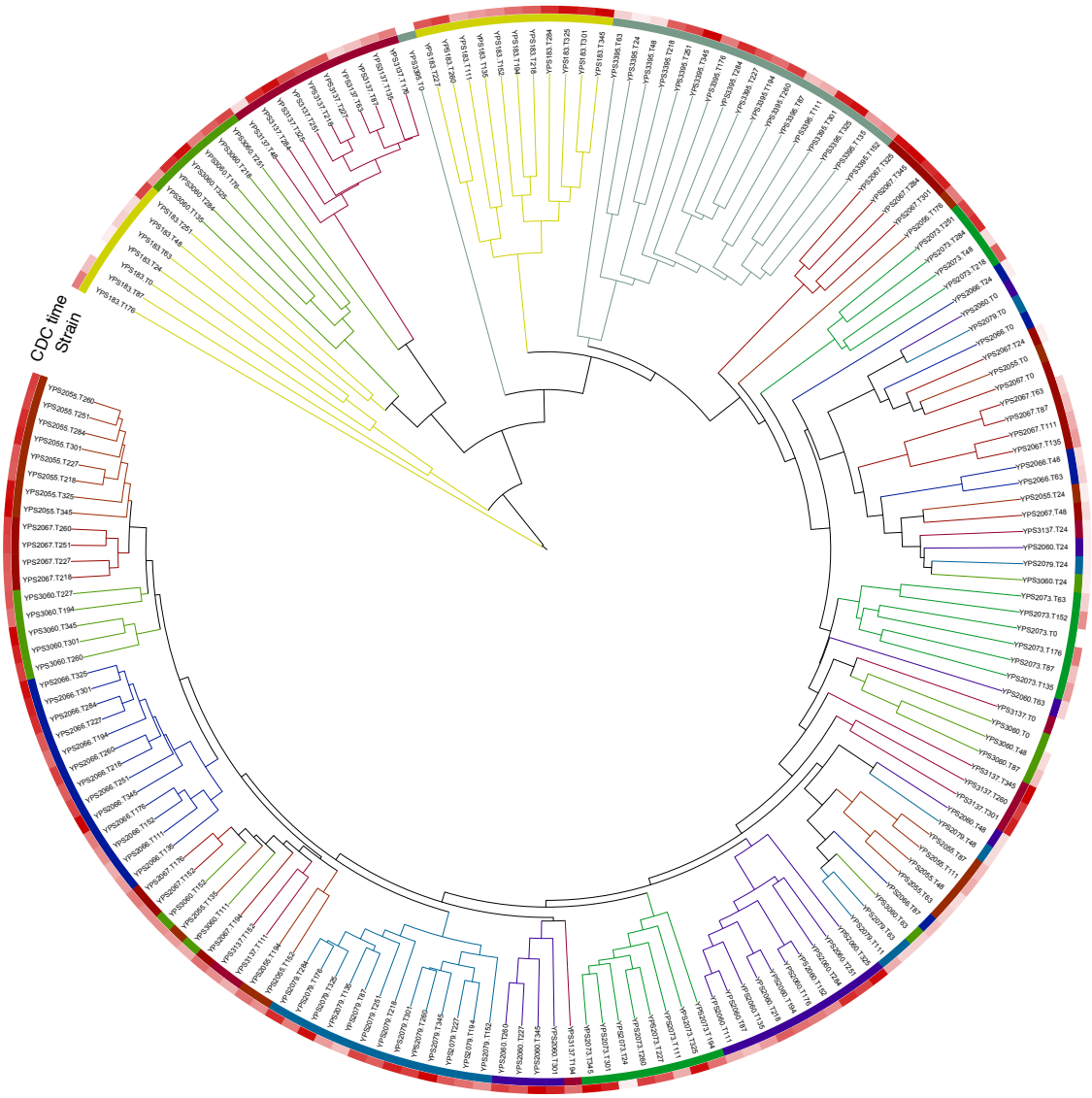


Figure 3.4: Hierarchical clustering of the entire CDC data set

The average linkage method was used together with the Pearson correlation similarity metric. Branch and inner-ring colors identify each strain, while outer-ring colors indicate progression through the CDC as a percentage (0: 0 min.→white through 100: 345 min.→red). The data set consists of genome-wide expression levels at 18 timepoints for 10 strains.

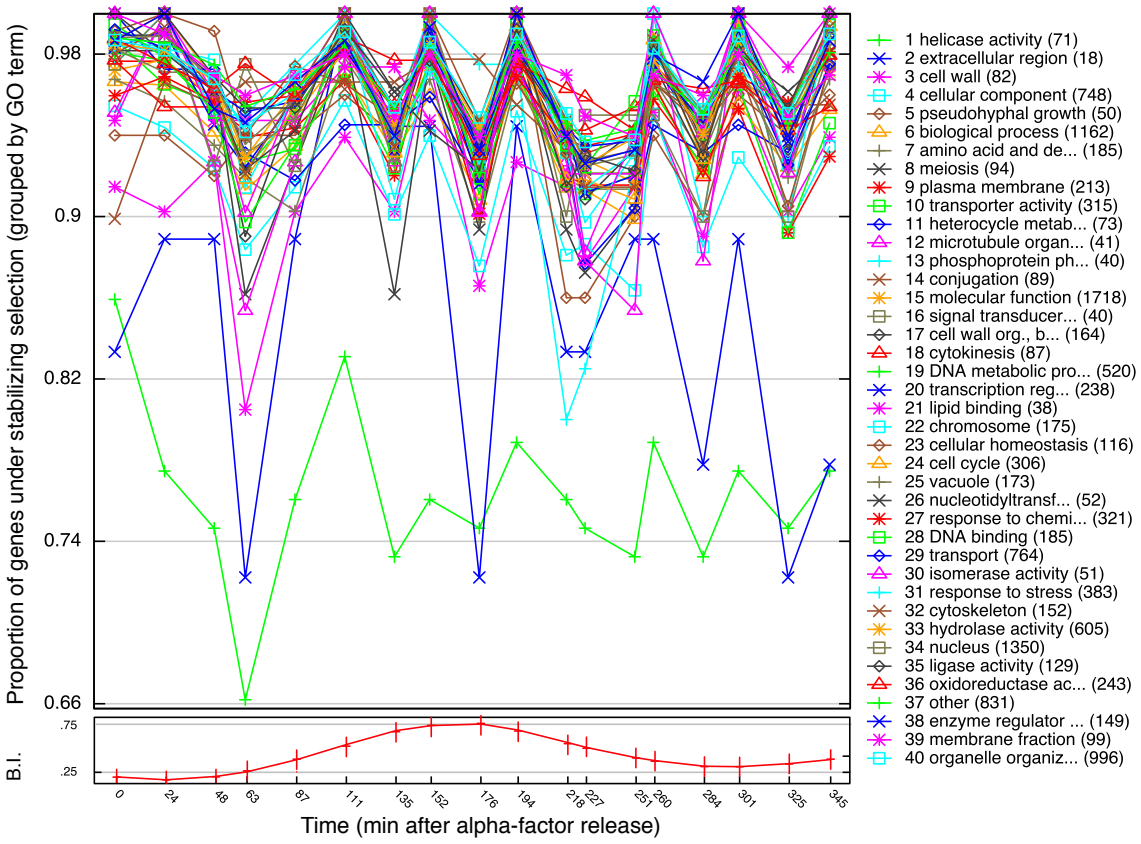


Figure 3.5: Profiles for the 40 GO-Slim terms which exhibit the lowest average proportion of genes under stabilizing selection

Profiles are plotted across the CDC and ranked by the average proportion of genes under stabilizing selection. Numbers in parentheses following each label indicate the number of genes with that label. The bottom panel illustrates the average *S. cerevisiae* budding index (BI) profile.

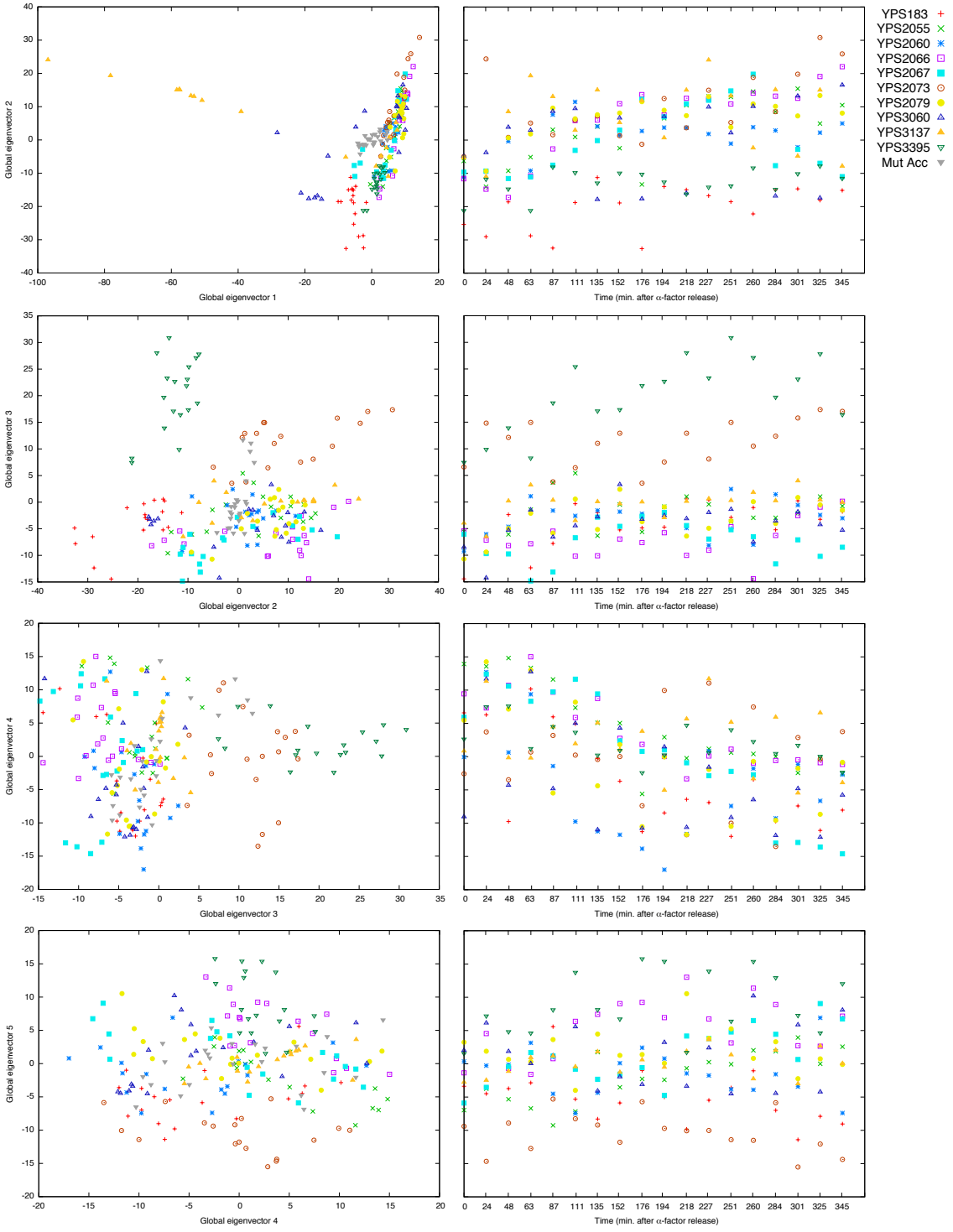


Figure 3.6: Visualization of CDC-transcriptomes using 2-dimensional SVD projections

(Left) 2-dimensional SVD projections of the entire CDC data set (18 timepoints for 10 strains) as well as mutation accumulation data. Data were mean centered and projected onto pairs of the top 4 global eigenvectors: (1^{st}, 2^{nd}), (2^{nd}, 3^{rd}), (3^{rd}, 4^{th}), and (4^{th}, 5^{th}). **(Right)** Plots of corresponding y-axis values (left) indexed by CDC timepoints. Colors group data points by strain. The first global eigenvector is defined by extreme variation in YPS3137, while global eigenvectors 2 and 3 appear to capture trajectories of variation shared by all strains. Eigenvectors 2 and 3 may thus comprise the primary CDC progression axes. Notably the *S. paradoxus* isolate appears to exhibit the most divergent trajectory (second row, left and right).

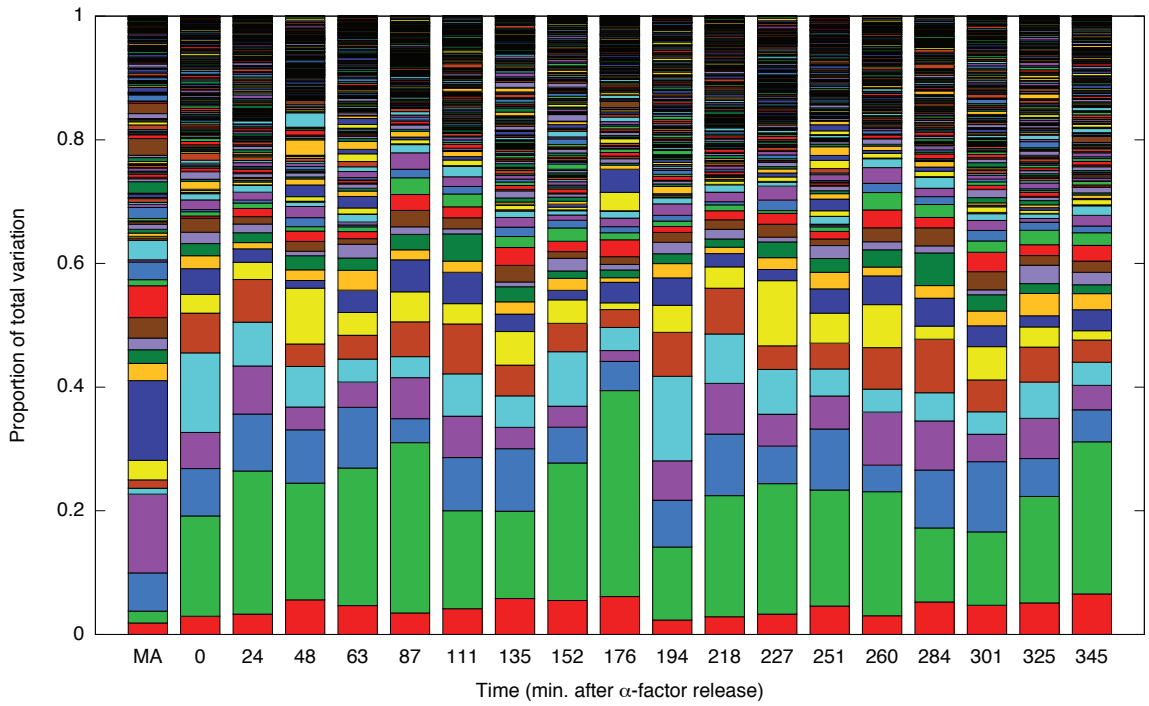


Figure 3.7: Comparison of cumulative eigenvalue distributions for MA line data (23 samples) and for *S. cerevisiae* CDC data at each timepoint (9 samples each)

All data were re-coordinatized by projection onto a common basis of the 162 global eigenvectors derived from SVD of the entire *S. cerevisiae* CDC data set. Each eigenvalue distribution illustrates the proportion of expression variation explained by each global eigenvector.

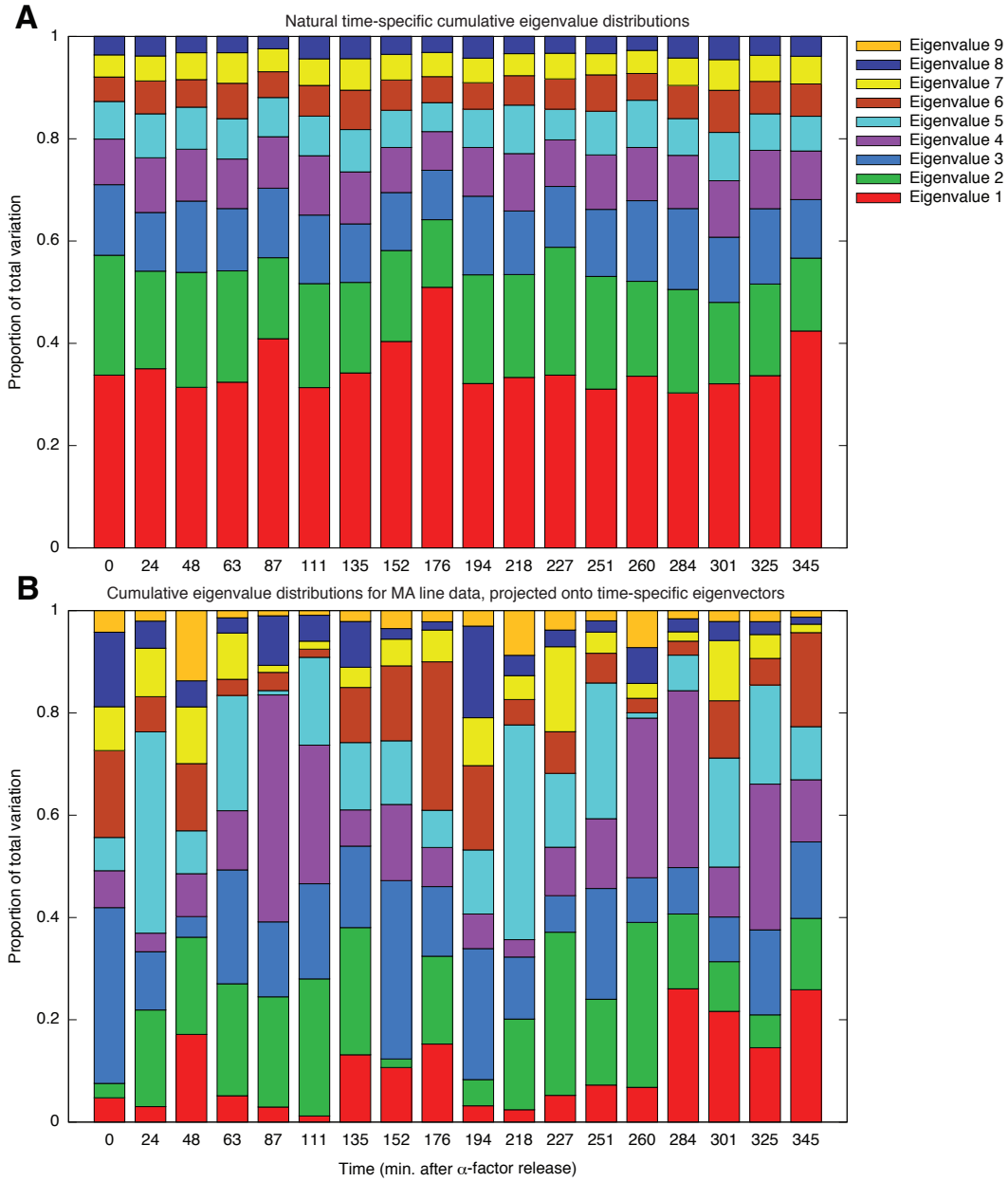


Figure 3.8: Cumulative eigenvalue distributions for natural CDC-transcriptomes and MA lines

Cumulative distributions **(A)** for eigenvalues of natural time-specific eigenvectors (CDC-directions) and **(B)** for eigenvalues of MA line data projected onto the natural time-specific eigenvectors. Eigenvalues were obtained by independent SVD of mean-centered expression data (9 *S. cerevisiae* samples) at each of the 18 timepoints. Stacked bars indicate the proportion of total variation explained by a particular eigenvector.

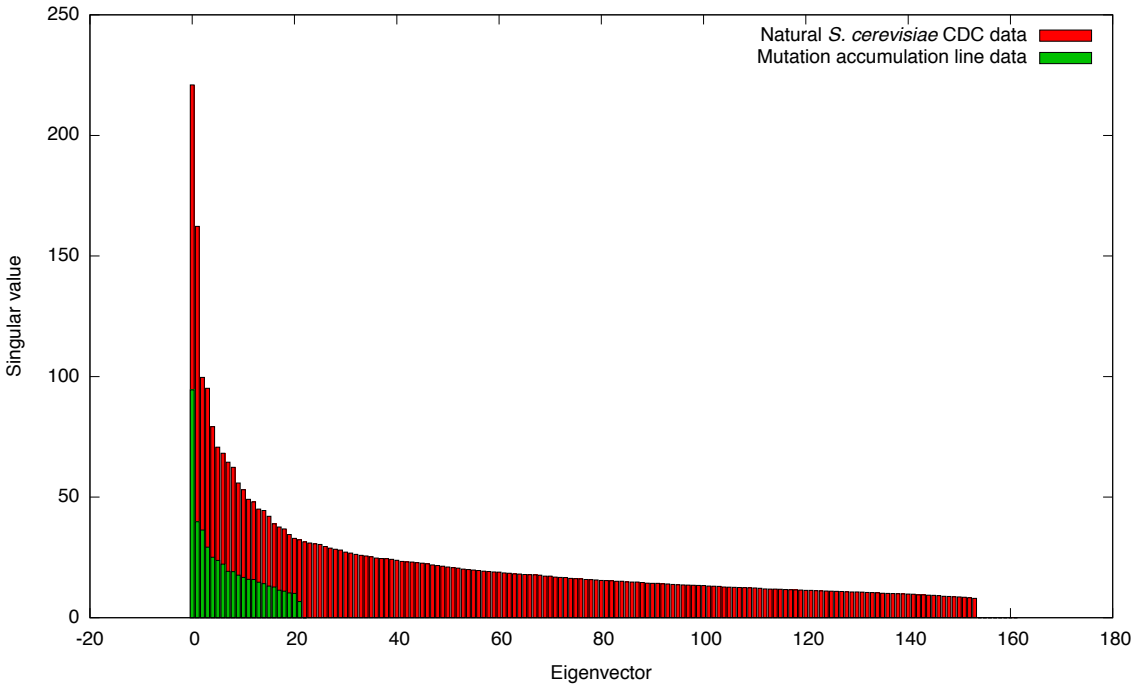


Figure 3.9: Comparison of singular value distributions for the *S. cerevisiae* CDC data (162 samples) and MA line data (23 samples)

Singular values were obtained by SVD of each data set after mean centering. The overlay illustrates that singular values estimated from time-series data greatly exceed those estimated from unsynchronized (time-averaged) data.

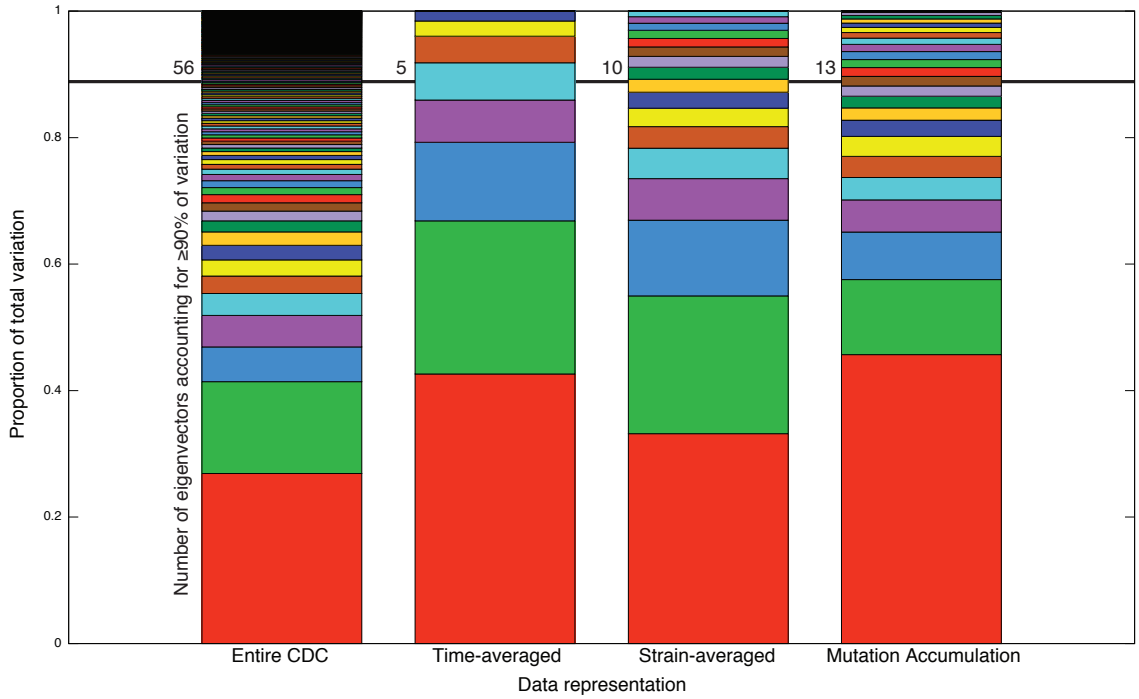


Figure 3.10: Comparison of transcriptome cumulative eigenvalue distributions

From left to right: the entire CDC data (162 samples), time-averaging of the entire CDC data (9 samples), strain-averaging of the entire CDC data (18 samples), and MA line data (23 samples). Eigenvalues were obtained by SVD of each data set after mean centering. The number of eigenvectors required to explain at least 90% of the variation in each data set are 56, 5, 10, and 13, respectively.

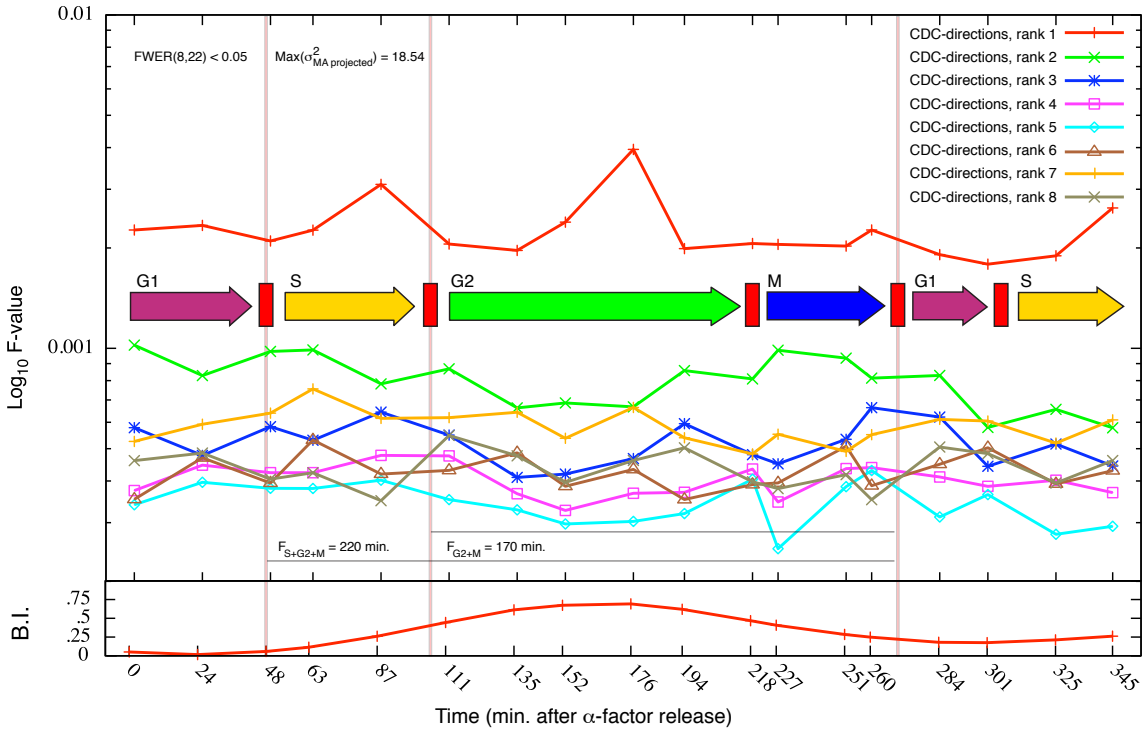


Figure 3.11: Multivariate evolutionary transcriptome variability across the CDC for each of the top 8 CDC-directions

F-values were computed for each CDC-direction as the ratio of per-generation natural to neutral multivariate variances, scaled by degrees of freedom (8 and 22, respectively). Natural variances were computed by projecting each *S. cerevisiae* strain's expression data onto the respective CDC-direction at each timepoint. Neutral variances were computed by projecting MA expression data onto the same CDC-directions and then averaging over the 18 variance values for a given rank (since MA expression data are effectively time-averaged). Each average neutral variance was used to calibrate the 18 natural variances at that rank. The 9^{th} curve (not shown) appears similar to curves 2–8 but has an average F-value of 2.9×10^{-27} as there are only 8 degrees of freedom. Vertical lines indicate approximate incidence of CDC-phase transitions: G_1/S: 47 min., S/G_2: 97 min., M/G_1: 267 min. These were staged using a calibrated CDC period of 267 min., previously reported intervals $F_{S+G_2+M} = 220$ min. and $F_{G_2+M} = 170$ min. (*34*), and CDC-expression data (for G_2/M; see Figure 3.34). The bottom panel shows the *S. cerevisiae* budding index profile averaged over strains.

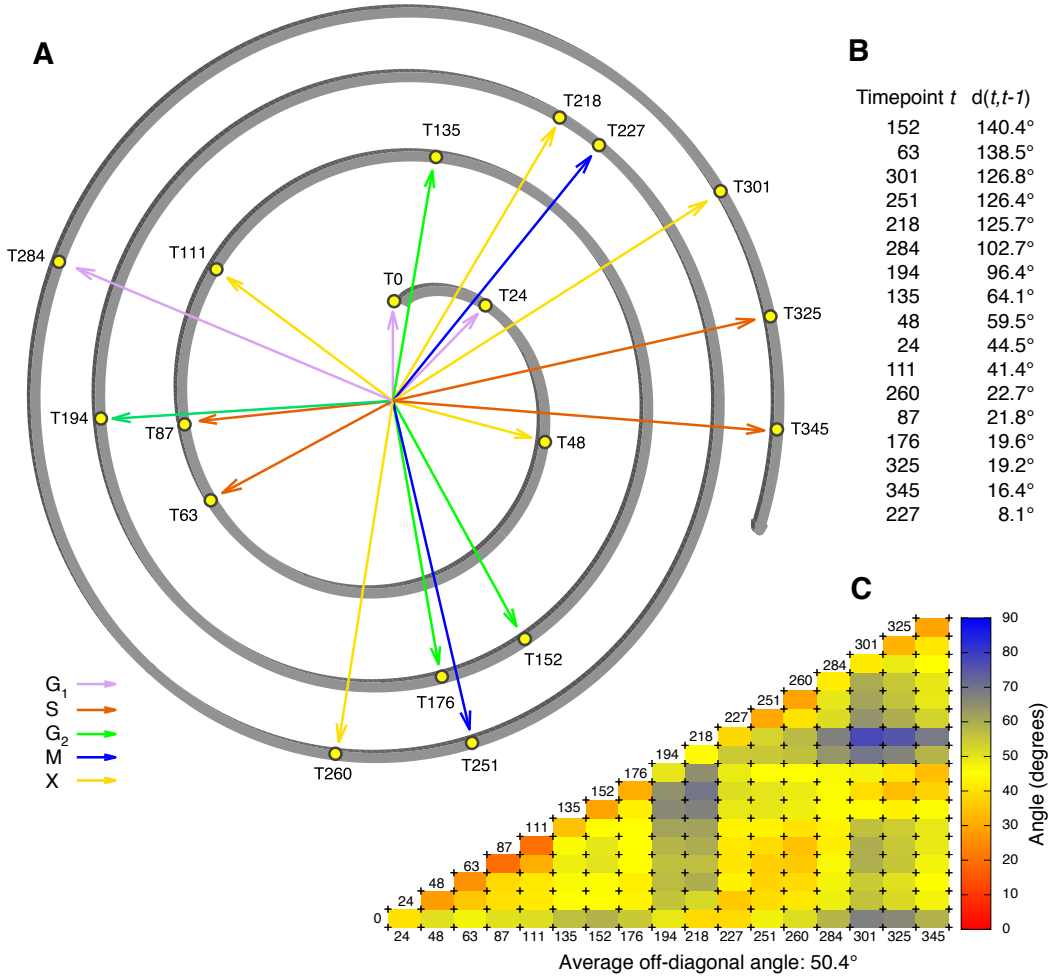


Figure 3.12: Temporal variability in CDC-transcriptome evolutionary covariance structure. **(A)** Spiral 2D projection showing angles between major directions of covariation at successive timepoints. Arrow colors indicate approximate CDC-phase. Xs denote CDC-phase transitions. Vector lengths are arbitrary. **(B)** Successive angles from (A) ranked by magnitude of change. **(C)** Heat map of angular changes in the major direction of covariation between all pairs of timepoints.

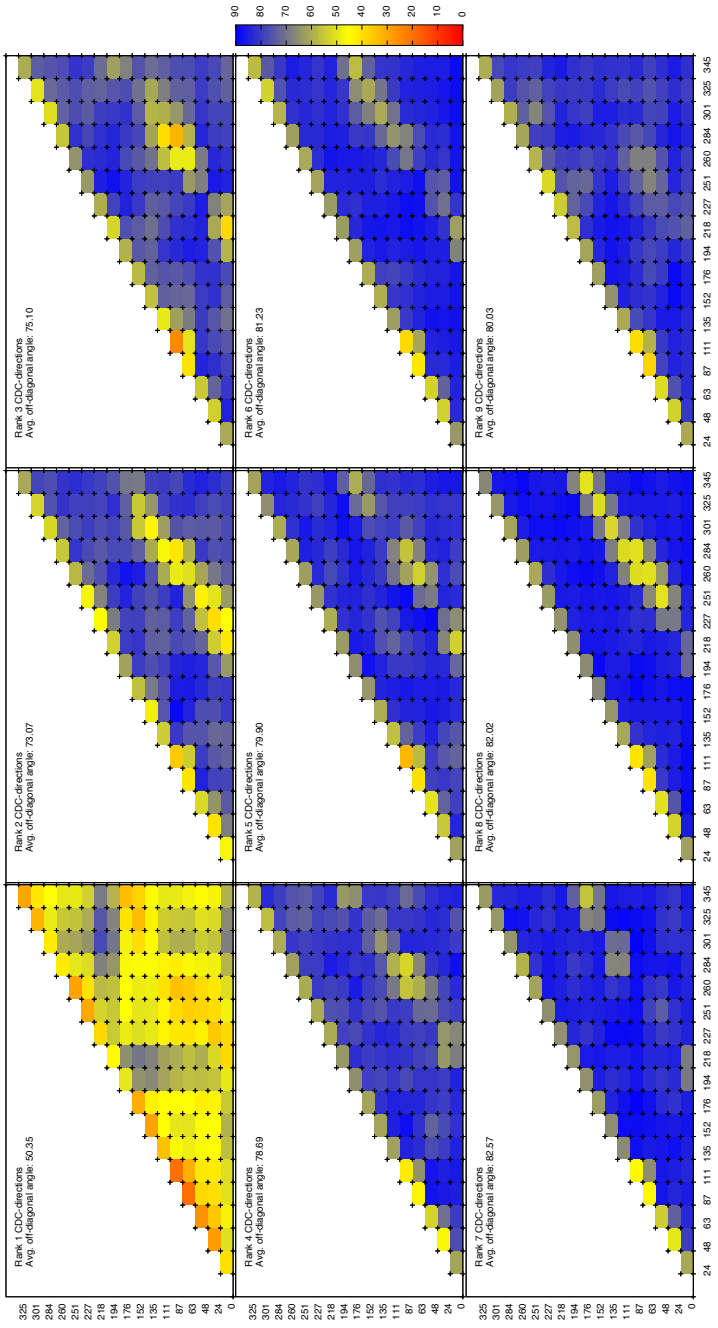


Figure 3.13: Angular distance matrices relating pairs of CDC-directions of the same rank

Axes indicate the 18 sampled CDC timepoints. Only unique pairs of timepoints are shown. The off-diagonal average angle of each matrix is indicated within each panel. The top left panel is repeated from Figure 3.12C and corresponds to the major (rank-1) directions of variation.

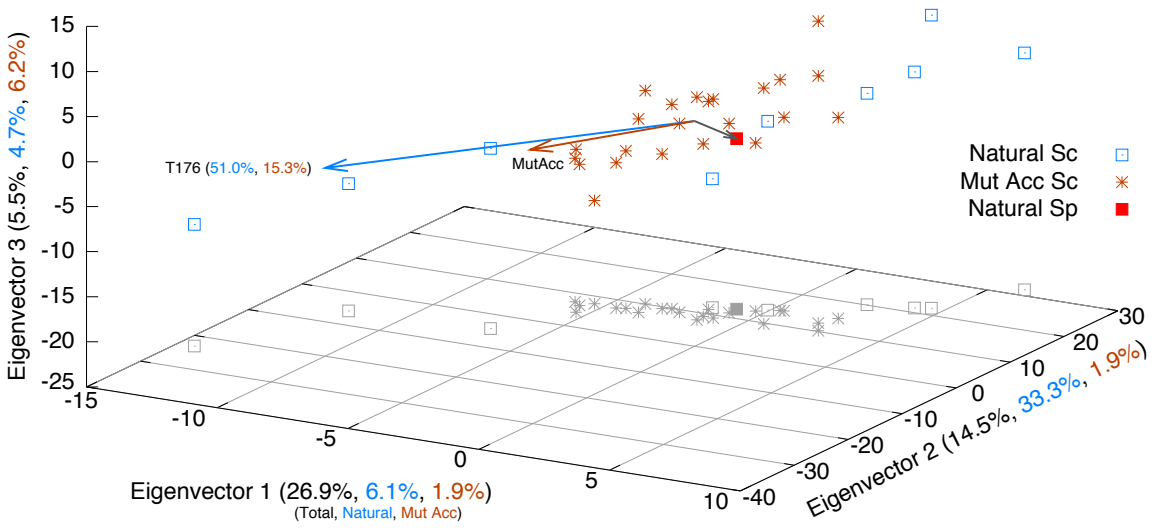


Figure 3.14: Example projection of expression data onto the top three global eigenvectors of *S. cerevisiae*

CDC-expression at 176 min. is shown for all *S. cerevisiae* and *S. paradoxus* as well as unsynchronized expression for the 23 MA lines. Percentages of total variation explained by each of the top 3 global eigenvectors are reported below each axis and correspond to entire CDC data (black), *S. cerevisiae* data at 176 min. (blue), and MA line data (brown). Percentages of total variation explained by the major CDC-direction at 176 min (blue arrow). are also shown for *S. cerevisiae* data at this timepoint (blue) and MA line data (brown). The label 'T176' indicates the major CDC-direction of variation among *S. cerevisiae* strains at 176 min. The brown arrow labeled 'Mut Acc' indicates the direction of largest neutral variation. The gray arrow indicates displacement vector direction for *S. paradoxus* at 176 min. Note that the arrows are meant as illustrations of eigenvectors that are inherently undirected. Gray shapes show 2D projection of the 3D data points on the top 2 global eigenvectors.

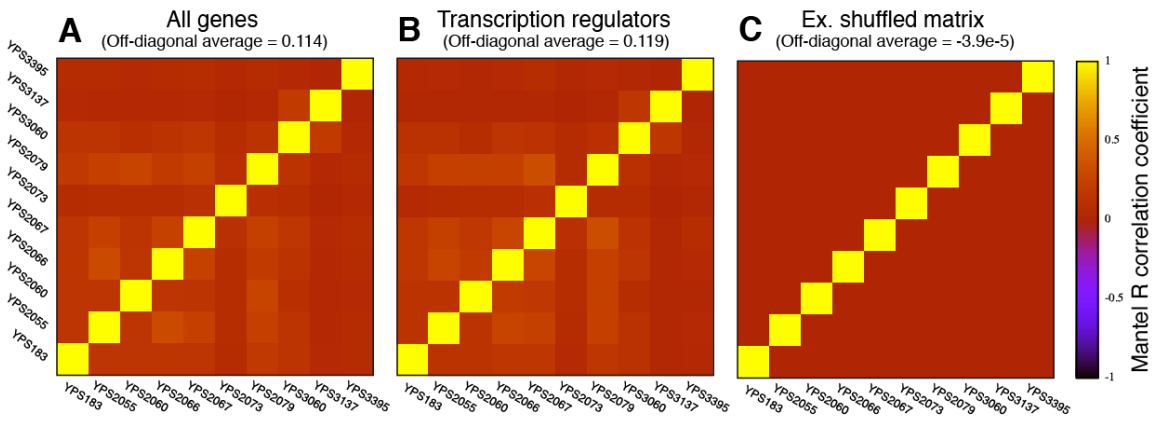


Figure 3.15: Heat maps illustrating Mantel matrix correlations between CDC-expression correlation matrices of natural strains

Each 10×10 heat map illustrates correlations using **(A)** all genes ($n = 6082$), **(B)** transcriptional regulators ($n = 266$), or **(C)** row and column shuffled genome-wide expression data ($n = 6082$). Each cell indicates the value of Mantel's *R* matrix correlation coefficient between a pair of strains. The correlation coefficient between the genome-wide and TF heat maps is 0.95 ($P < 0.001$).

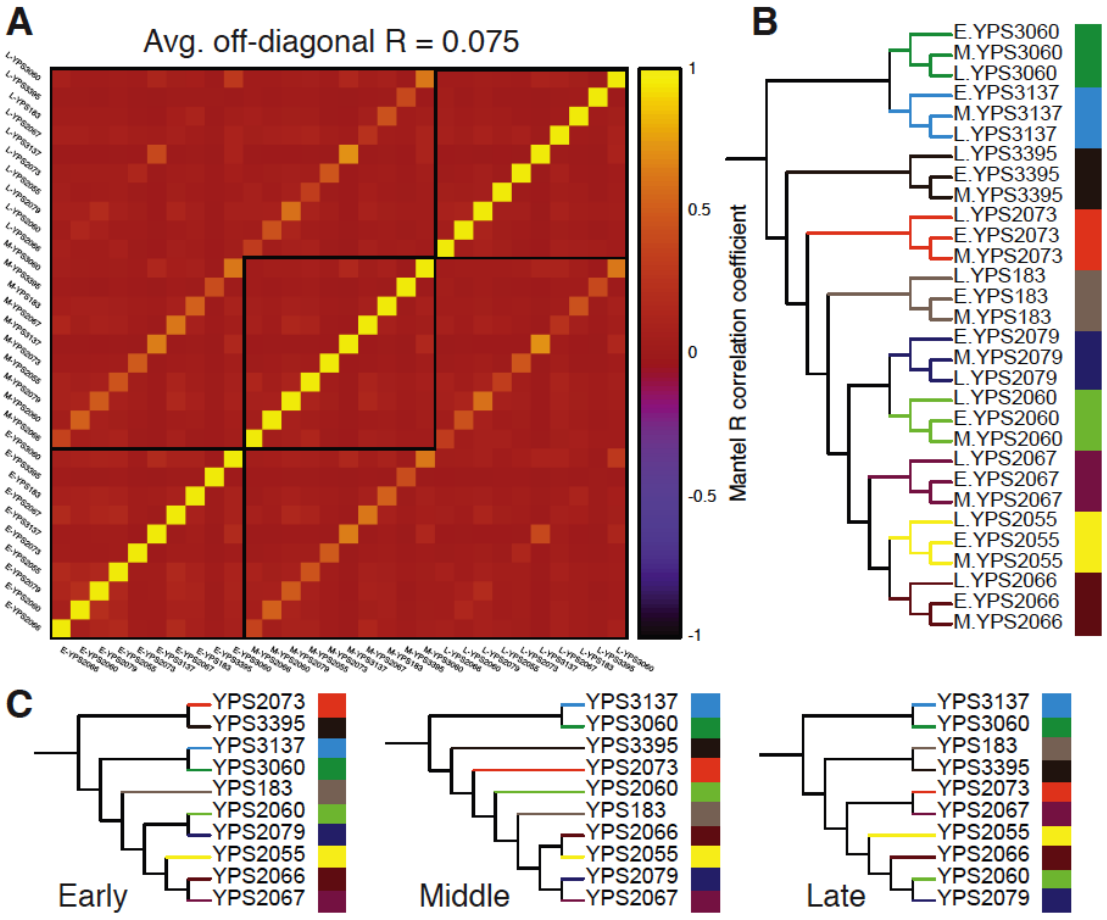


Figure 3.16: Evolutionary divergence of CDC-transcriptome coexpression structure within and between CDC-phase groups

(A) Heat map of Mantel matrix correlation coefficients between pairs of strains for each of 3 CDC-phase groups (Early: E, Middle: M, Late: L), corresponding to the first, middle, and last 9 sampled timepoints. Correlations were computed between pairs of 6082×6082 genome-wide CDC-expression correlation matrices. **(B)** Hierarchical clustering of the correlation matrix shown in (A). **(C)** Hierarchical clusterings for data within each CDC-phase group, corresponding to the 3 main diagonal blocks (outlined in (A)). Clustering was performed using average linkage with the Pearson correlation metric.

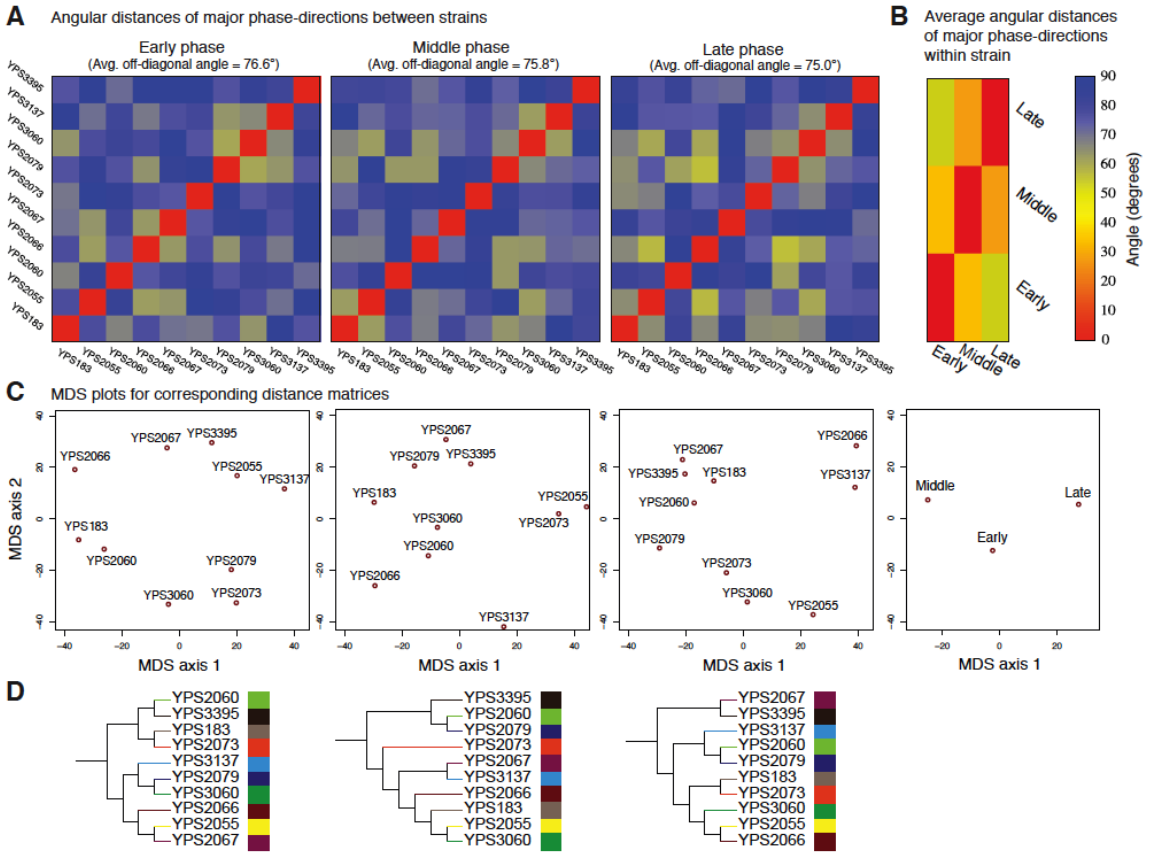


Figure 3.17: Evolutionary divergence of major directions of covariation within CDC-phase groups

(A) Heat maps of angular distances between pairs of major phase-directions for 3 CDC-phase groups (Early: E, Middle: M, Late: L), corresponding to the first, middle, and last 9 sampled timepoints. Each phase-direction corresponds to the major direction of covariation within a phase group, identified by SVD. **(B)** Heat map of angles relating the major phase-directions across the CDC-phase groups of each strain. Cell color indicates the average angle across 10 strains. Significance of all angles in (A) and (B) was established for each heat map using a random angles test (FWER < 0.05). **(C)** Classical metric multidimensional scaling (MDS) plots for each phase, derived from corresponding angular distance matrices. **(D)** Hierarchical clusterings of the angular distance matrices, performed using average linkage with the Pearson correlation metric.

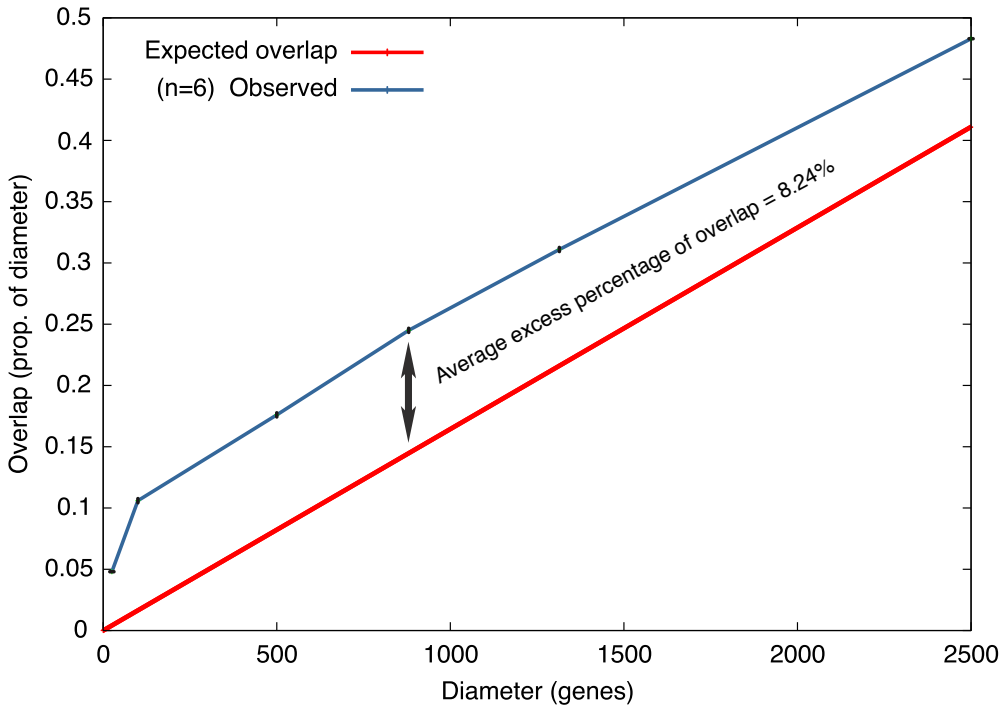


Figure 3.18: Expected and observed proportions of k-module overlap for increasing numbers of genes k

The expected k-overlap proportion is equal to the module size k scaled by the number of genes ($n = 6082$). The average excess in percentage of overlap compared to random expectation is 8.24%.

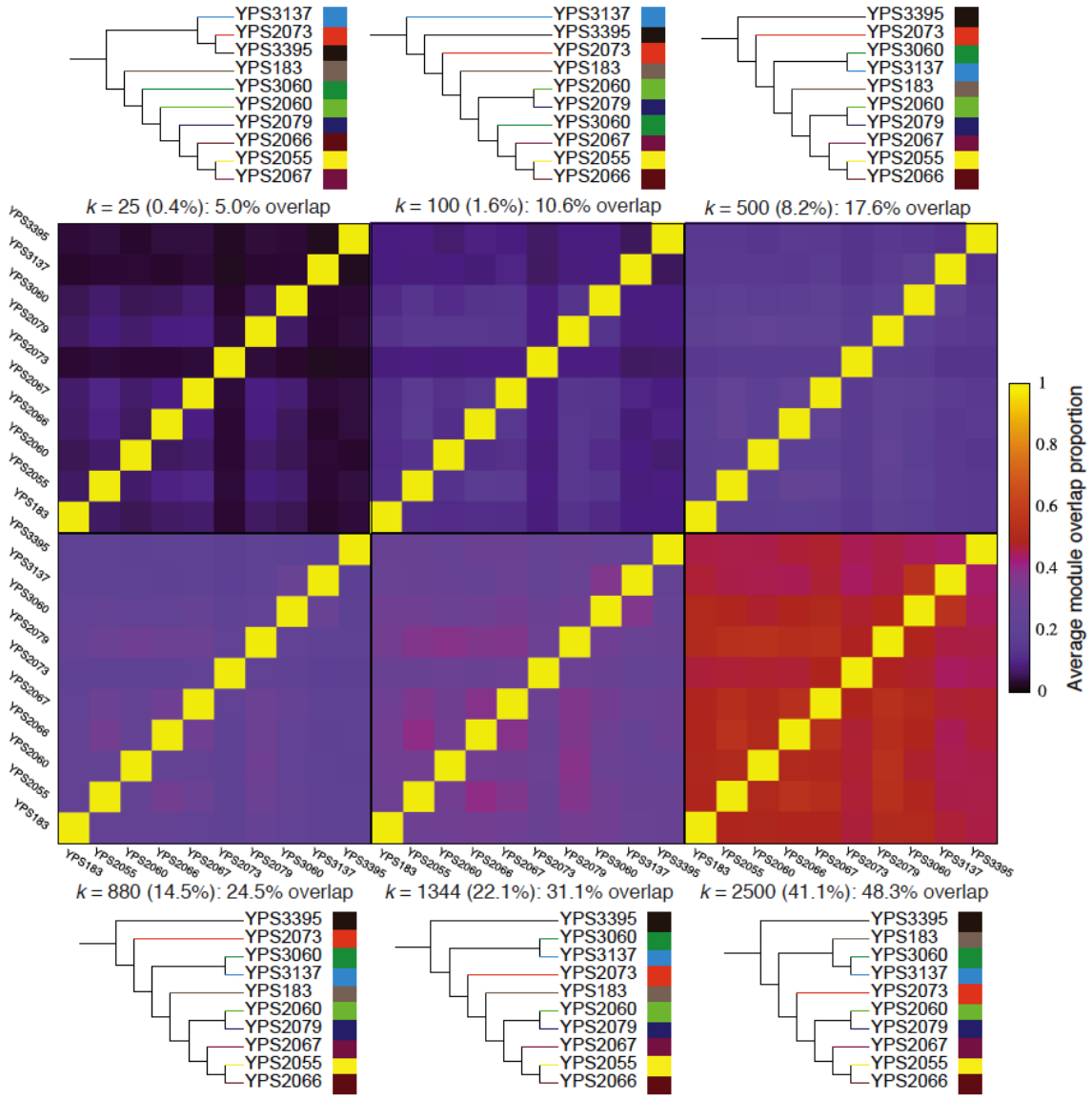


Figure 3.19: Heat maps and hierarchical clusterings indicating proportions of overlap between gene modules across strains

A module is defined for every gene as the set of its k top correlating genes. Module diameters k include 25, 100, 500, 880, 1344, and 2500 genes. Cell colors reflect the proportion of overlap of the modules between strains, averaged over all significant genes ($P < 1/250$). Significance was assessed for each module comparison (every gene between strains) by permuting the expression values of all genes in one strain 250 times and computing a (null) distribution of overlap statistics. Hierarchical clusterings were performed using average linkage with the Pearson correlation metric.

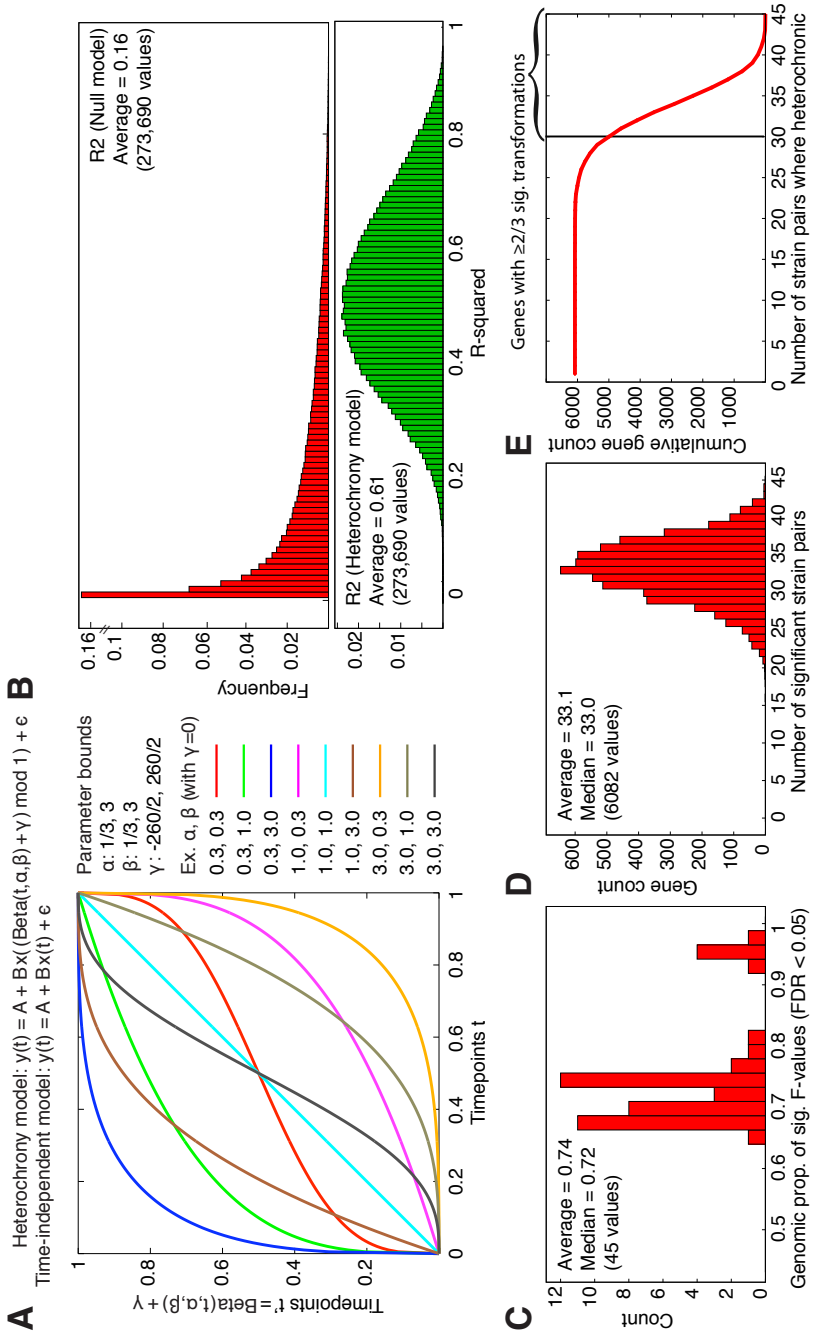


Figure 3.20: Results of the heterochrony regression model explaining time-dependent changes in gene expression trajectories between strains

The model was fit to single period, Z-standardized CDC-expression data for each gene given a query strain and a target strain. With 10 strains, there are 45 query-target strain pairs. **(A)** Formulation of the time-independent (null) and heterochrony (alternative) regression models. The heterochrony model involves a mapping of timepoints between strains, estimated using the Beta cumulative distribution function, which generates smooth, continuous, and invertible transformations on $[0, 1]$. Since the phase of a gene's expression may also change over evolution, a phase parameter γ was included. Transformed timepoints were modulated around 1, so that transformations are defined with respect to a single CDC. Estimates of α, β, and γ were bounded within $[1/3, 3]$, $[1/3, 3]$, and $[-260/2, 260/2]$, where 260 is the CDC period. The light blue line ($\alpha = 1, \beta = 1, \gamma = 0$) describes the null model time transformation, where $t = t' = Beta(t, 1, 1) + 0$. **(B)** Distributions of R-squared values for the time-independent **(top)** and heterochrony **(bottom)** models. Both models were estimated identically, except that parameter values for the null model were fixed at ($\alpha = 1, \beta = 1, \gamma = 0$). **(C)** Distribution of the proportion of significant F-values (genes) over the 45 strain comparisons (FDR < 0.05). **(D)** Distribution of the number of significant strain comparisons over genes. **(E)** Plot showing the number of genes significant in at least k comparisons versus k. A cutoff of $30/45 = 2/3$ was used to classify a subset of 4998 genes as heterochronic.

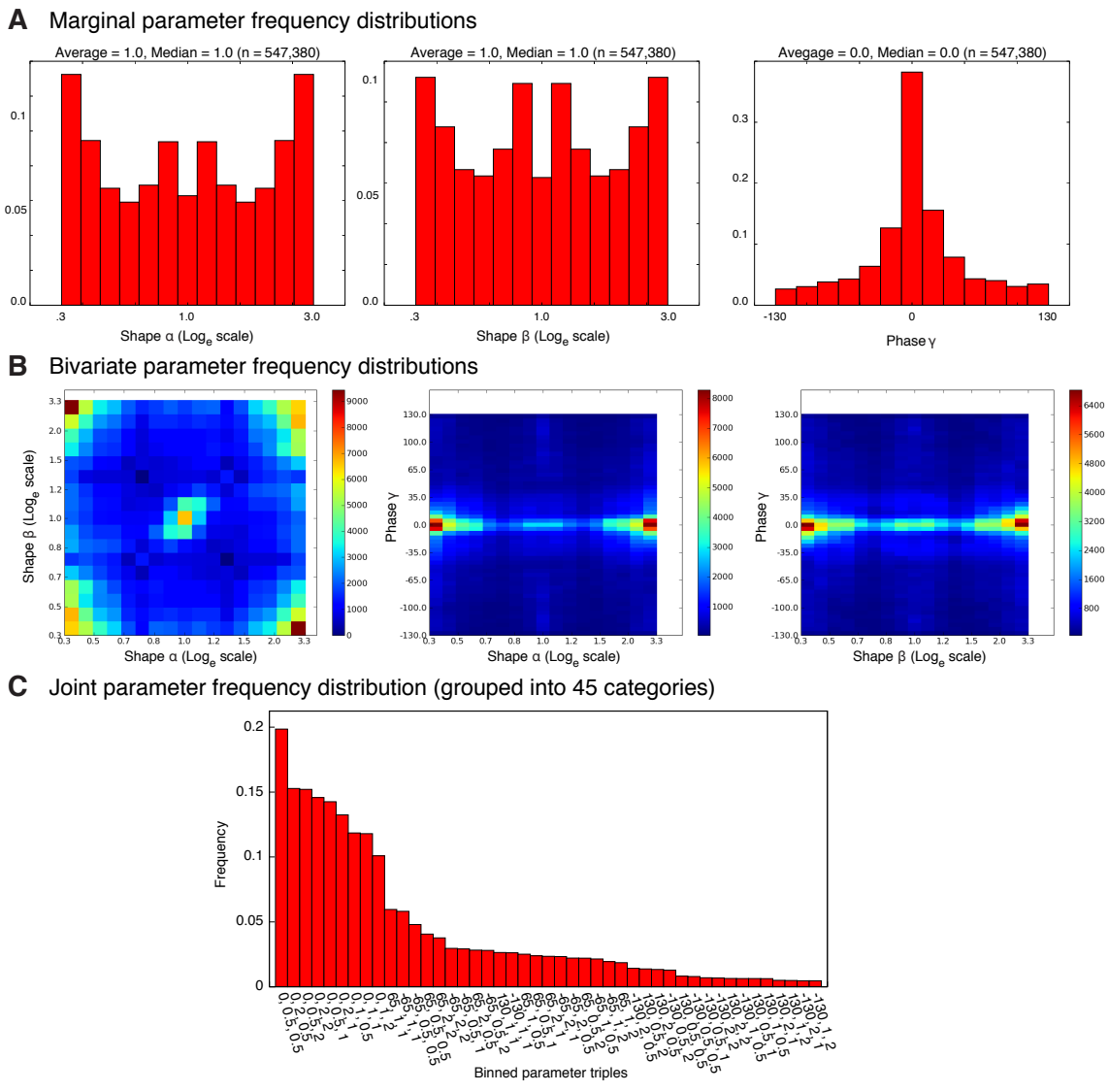


Figure 3.21: Frequency distributions of optimal time-domain parameters

(A) Marginal, **(B)** bivariate, and **(C)** joint frequency distributions for the set of optimal time-domain parameters estimated for 6082 genes i across all 45 unique pairwise strain comparisons j and their reciprocals (swapping dependent and independent variables) (6082×90 total parameter triples). The time-domain transformation model is invertible using the inverse of the time-domain parameters: $Y_i^j = f(Y_i^k|\alpha, \beta, \gamma)$ and $Y_i^k = f(Y_i^j|1/\alpha, 1/\beta, -\gamma)$. Thus the inverse of a shape parameter α is $-\alpha$ on the log scale. For visualization of the joint frequency distribution in (C) all parameter triples were first grouped into 45 categories before plotting.

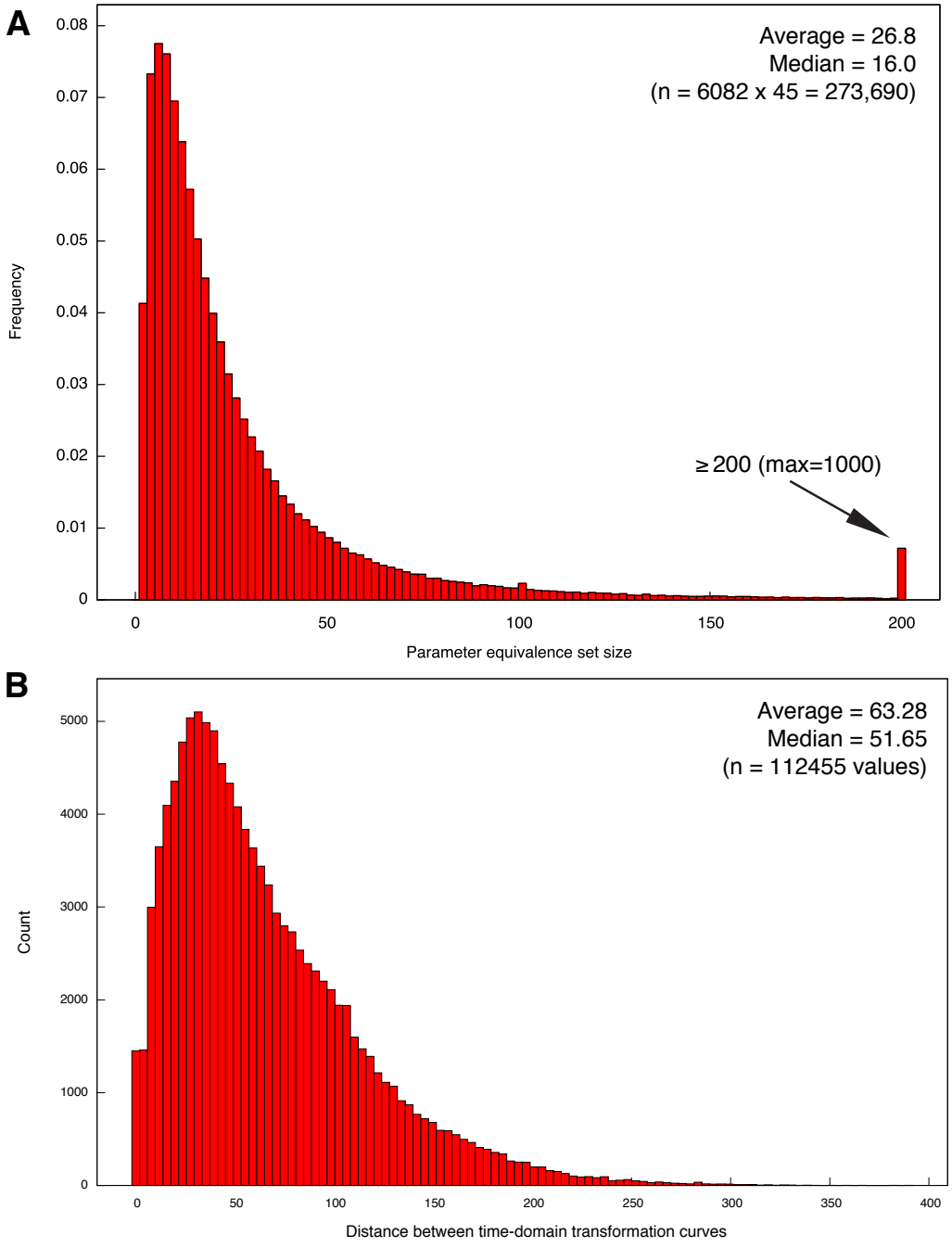


Figure 3.22: Distributions of 95%-equivalent timing curve ensemble sizes and distances between ensembles

(A) Distribution over the size of each gene's set of 95%-equivalent time-domain parameter triples, pooled over 45 strain comparisons. Each parameter triple defines a time-domain transformation curve (timing curve) between strains. Each equivalence set provides a 95% error bound on the estimate of gene's a timing curve between two strains. The last histogram bin shows the accumulation of set sizes greater than or equal to 200. **(B)** Sample distribution of distances computed between the equivalence sets of pairs of genes, pooled over strain comparisons. Distance between sets is the minimum root mean squared error over all pairs of timing curves.

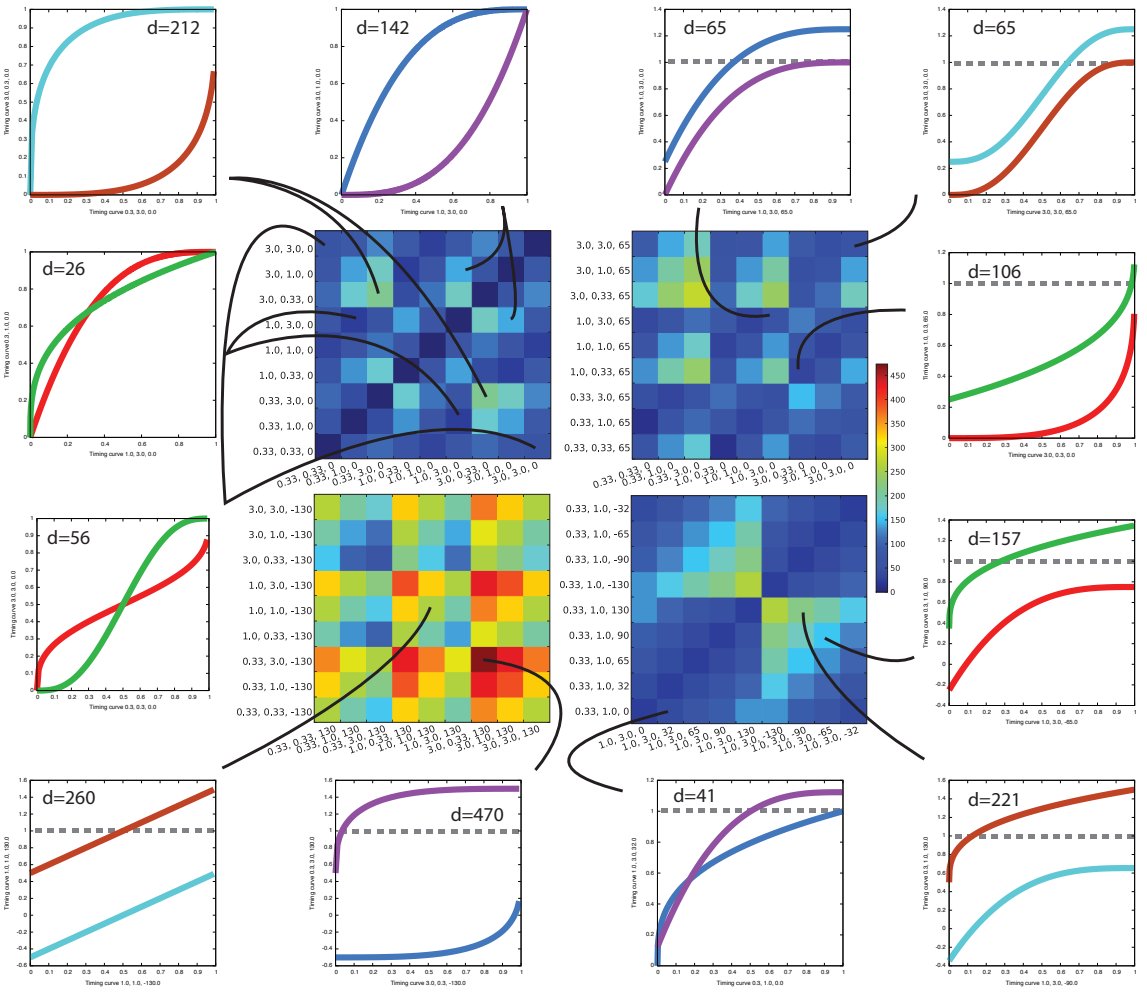


Figure 3.23: Heat maps and timepoint plots illustrating distances between pairs of timing curves

Timing curves are defined by parameter triples (α, β, γ), as shown: variation in α and β with $\gamma = 0$ **(top left)**; variation in α and β with $\gamma = 0$ vs. $\gamma = 65$ **(top right)**; variation in α and β with $\gamma = 130$ and $\gamma = -130$; and variation in γ with $(\alpha, \beta) = (1.0, 3.0)$ and $(\alpha, \beta) = (0.33, 1.0)$. Distance between a pair of triples was computed by root mean squared error, and is equal to the distance between the inverse of these triples. Each timepoint plot visualizes the timing curves defined by parameter triples as indicated. The distance metric integrates the total error between 2 timing curves across across the CDC. Gray dashed lines indicate the boundary of 1 CDC period; all curves that pass this line have some degree of phase offset ($\gamma \neq 0$).

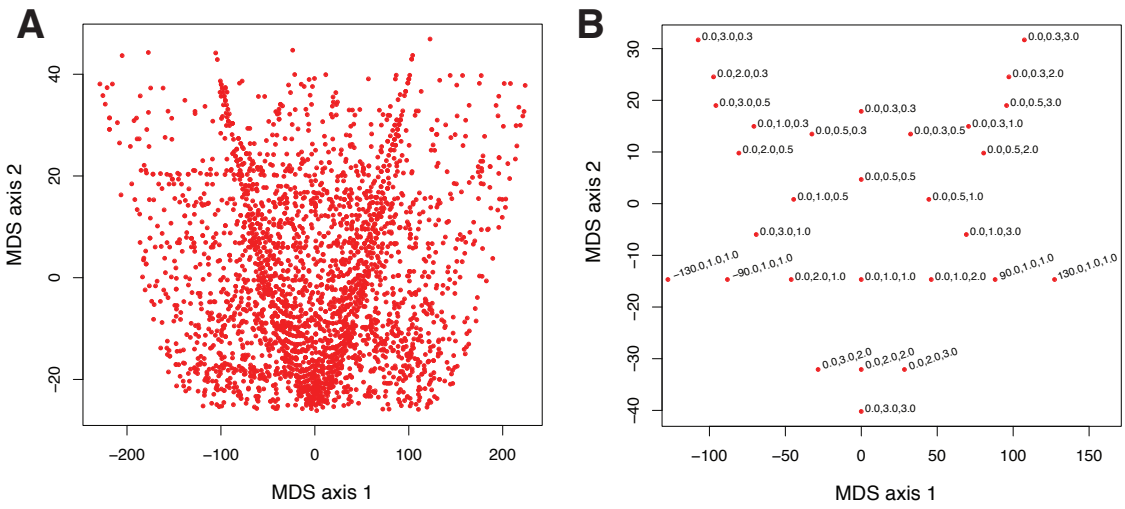


Figure 3.24: Visualization of timing curve distances by MDS

(A) Plot of 1000 randomly sampled time-domain parameter triples, visualized by metric multidimensional scaling (MDS) of the corresponding pairwise distance matrix in 2 dimensions. Distance between parameters was computed as RMSE between the timing curves induced by each parameter triple (see Figure 3.23 for examples). **(B)** MDS plot of 29 parameter triples shown with corresponding parameter labels. The 'V' shape in (A) and (B) is defined mostly by changes in α and β, while the broader scatter is defined by changes in all 3 parameters.

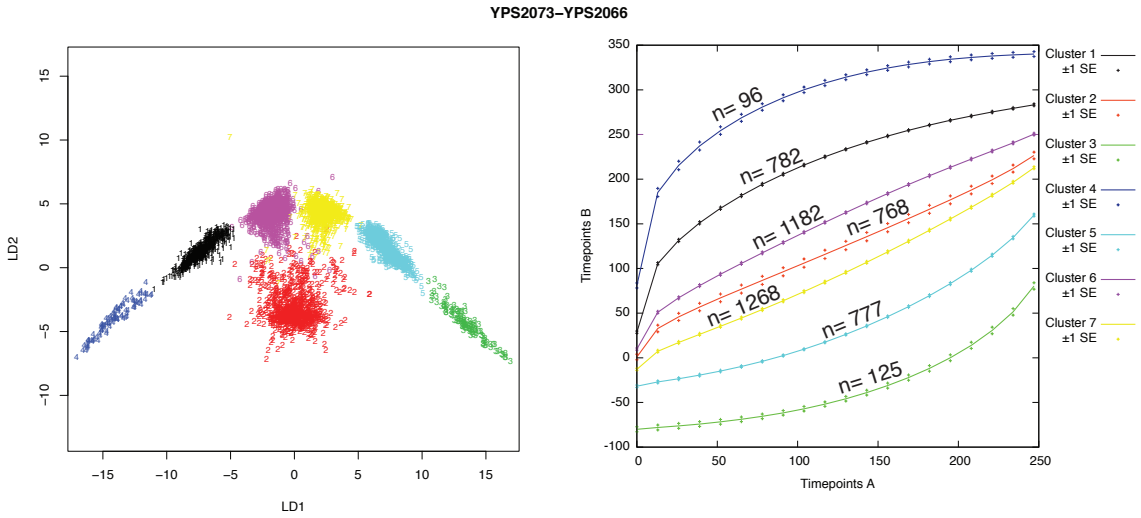


Figure 3.25: Linear discriminant distance matrix visualization and timing curves for each cluster.

(A) 2D linear discriminant analysis (LDA) plot of the distance relationships between 4998 heterochronic genes using parameters estimated from the comparison between YPS2073 and YPS2066. Distances between genes were computed as the minimum RMSE between all pairs of timing curves within each gene's 95%-equivalence timing curve ensemble. Point colors and numbers correspond to cluster labels for the genes, as determined by k-means clustering with $k = 7$ clusters and 100 random starts. **(B)** Timepoint plot showing the mean timing curves with 1 standard error for each of the clusters from (A). The number of genes in each cluster is indicated.

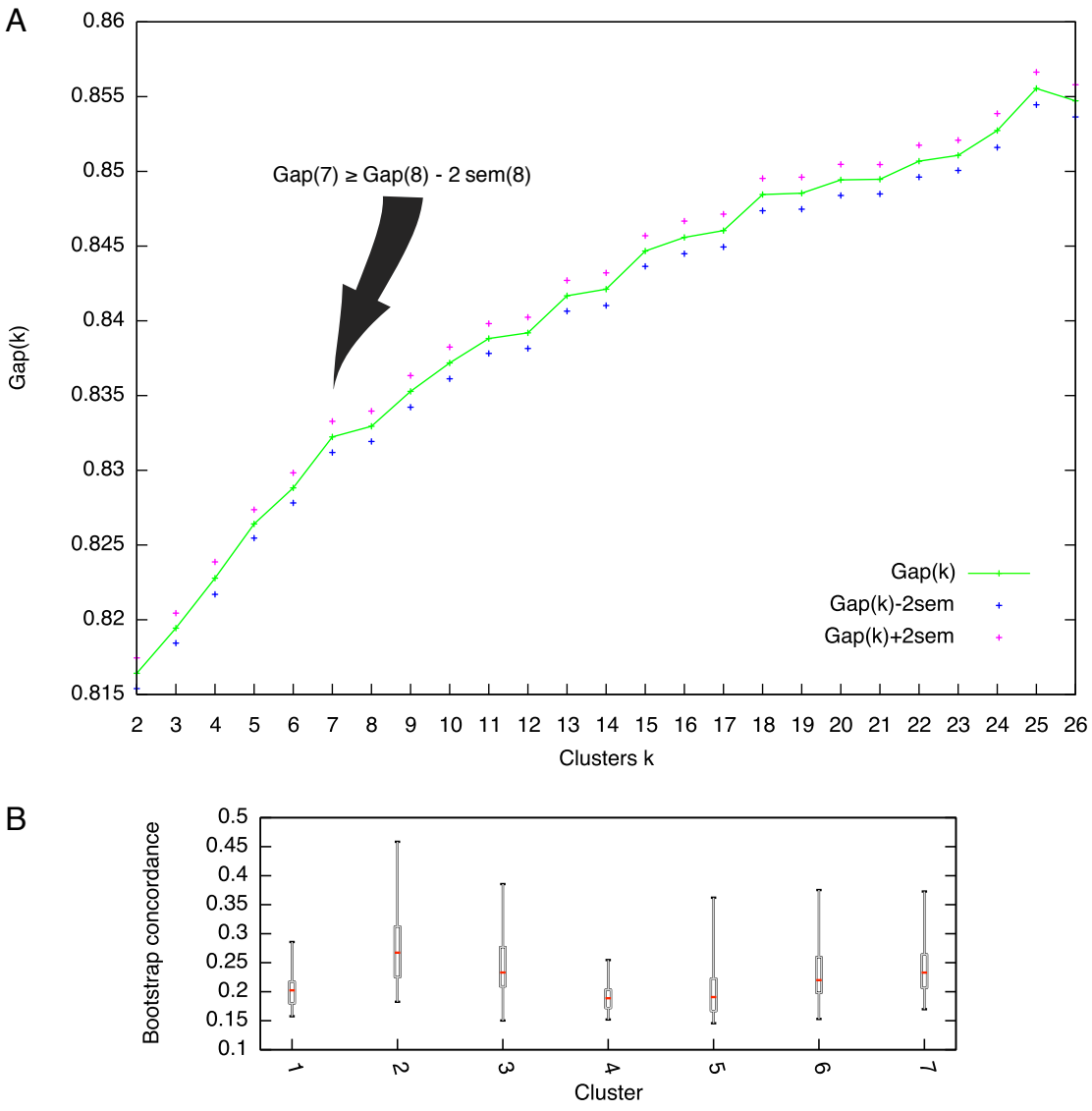


Figure 3.26: Gap statistic and bootstrap analyses of timing pattern clusters

(A) Gap statistic analysis, computed as the element-wise average of the 45 per-comparison timing pattern distance matrices. A gap statistic was computed for $2 \geq k \geq 26$, given the data matrix and the corresponding k-means clustering. Support for a particular k was identified as the smallest k such that $gap(k) \geq gap(k+1) - 2sem(k+1)$. In this way gap statistic analysis supports $k = 7$ clusters in the summary matrix. **(B)** Bootstrap analysis of the summary matrix clustering by randomly resampling 45 per-comparison distance matrices and recomputing a summary matrix and clustering it, 100 times. Boxplots show the distribution of cluster concordance between a bootstrapped clustering and the true clustering. Cluster concordance is the average concordance over the 7 clusters, where concordance is the fraction of genes in a true cluster which also appear in a bootstrapped cluster.

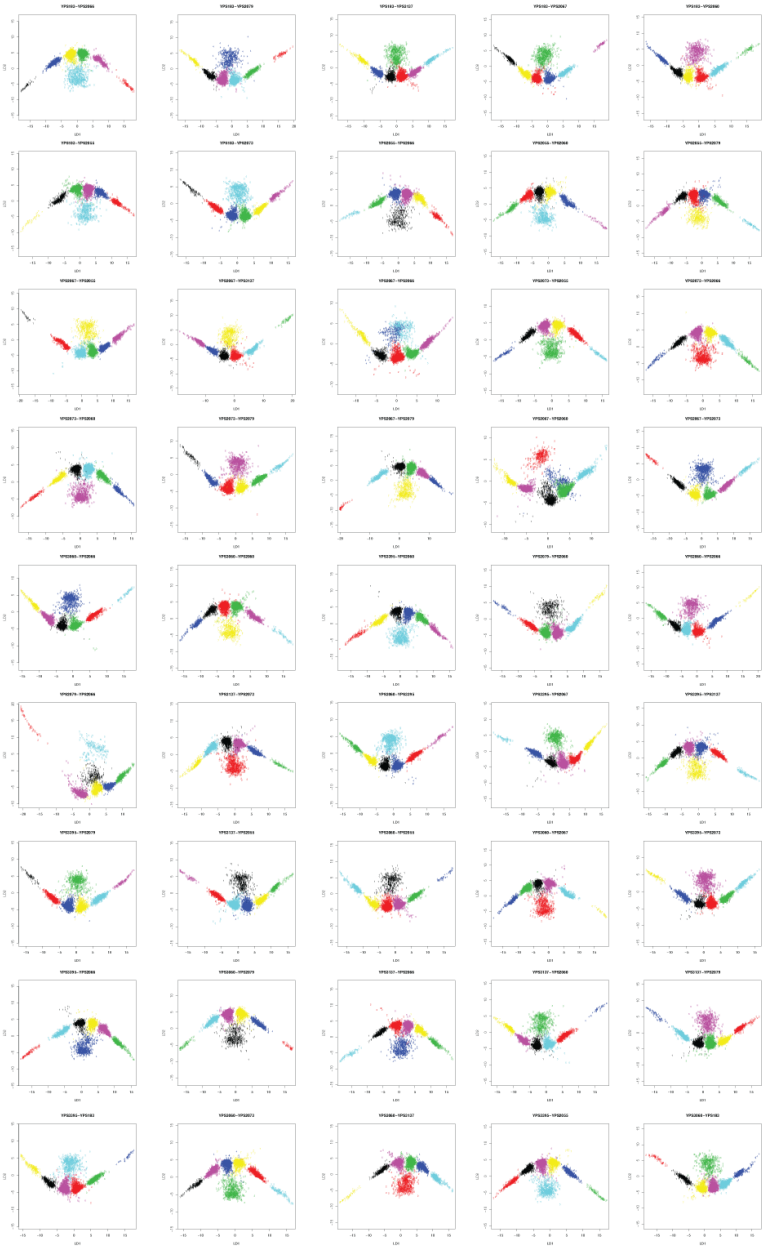


Figure 3.27: 2-dimensional linear discriminant plots of the distance relationships between timing curves for 4998 heterochronic genes for each of the 45 strain comparisons

Clusters were determined by k-means clustering with $k = 7$ clusters and 100 random starts.

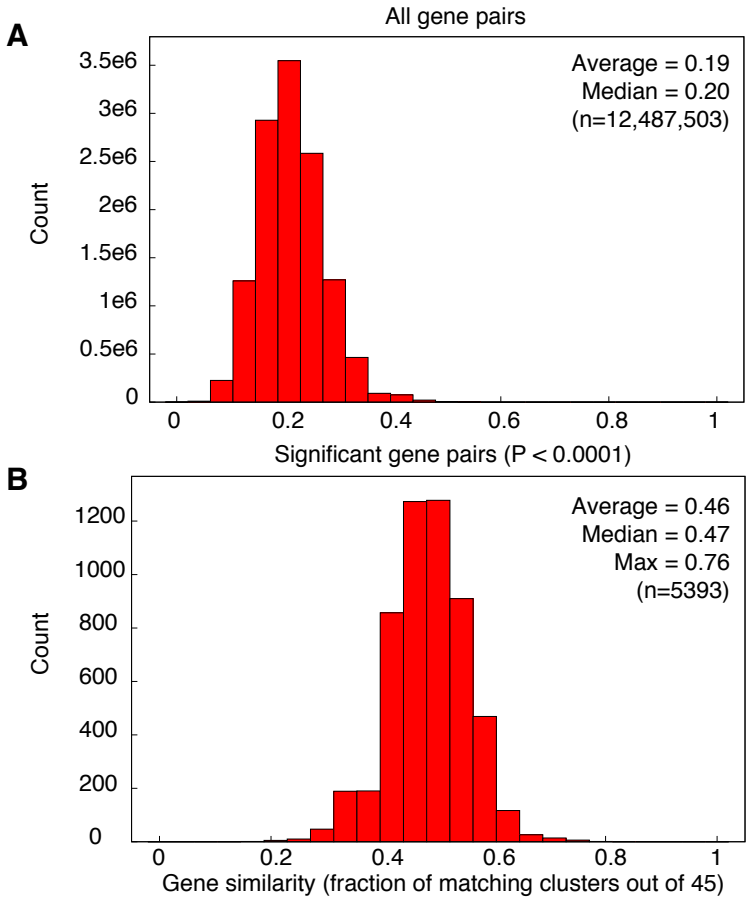


Figure 3.28: Distributions of co-cluster similarity between pairs of genes

Co-cluster similarity was determined using the length-45 cluster label profiles across strain comparisons for **(A)** all pairs of genes and **(B)** significant pairs of genes ($P < 10^{-4}$). Co-cluster similarity is the fraction of matching cluster labels between a pair of genes. P-values were computed using a binomial distribution with 45 trials and number of successes equal to the number of matching cluster labels. A success probability was computed separately for each strain comparison as the frequency of a cluster in that strain comparison scaled by the number of genes ($n = 4998$). The binomial probability was then computed as the binomial coefficient times the product of the success (or failure) probability of each strain comparison.

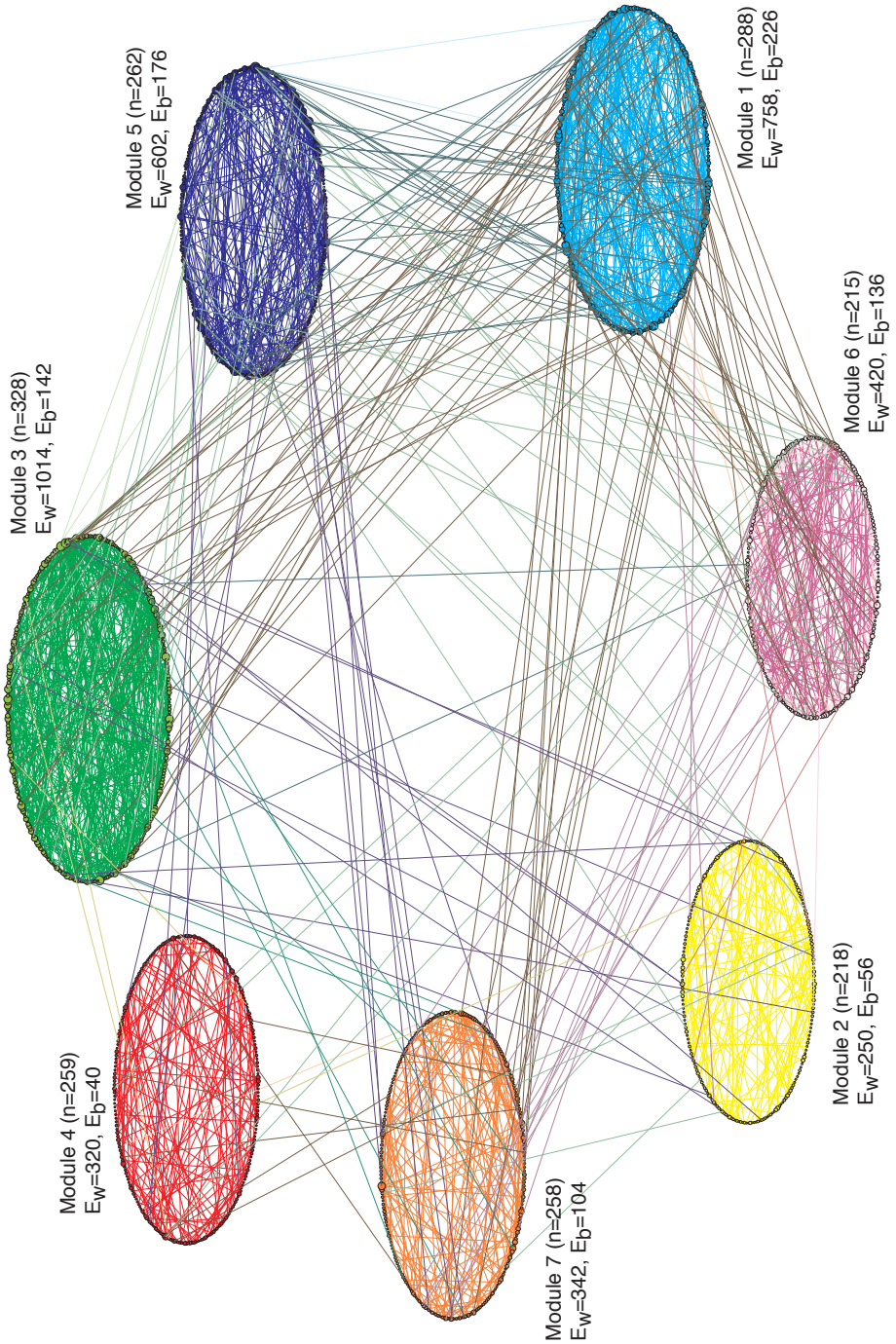


Figure 3.29: Modularity in the heterochronic gene interaction graph

This graph illustrates the significant heterochronic interactions between the 1828 genes closely associated with 7 timing pattern clusters ($P < 10^{-4}$). Clustering was obtained by k-means using 100 random starts. k was determined by gap statistic analysis (Figure 3.26A). Vertices (genes) are grouped and colored by cluster. Vertex size is proportional to the square root of its degree (total number of edges). Edges (interactions) are distinguished as within-cluster edges and between-cluster edges and are colored differently. The number of genes (n), within-edges (E_w), and between-edges (E_b) are shown for each cluster. The number of within-edges correlates with the number of genes in each cluster (Spearman's $r = 0.87, P = 0.005$).

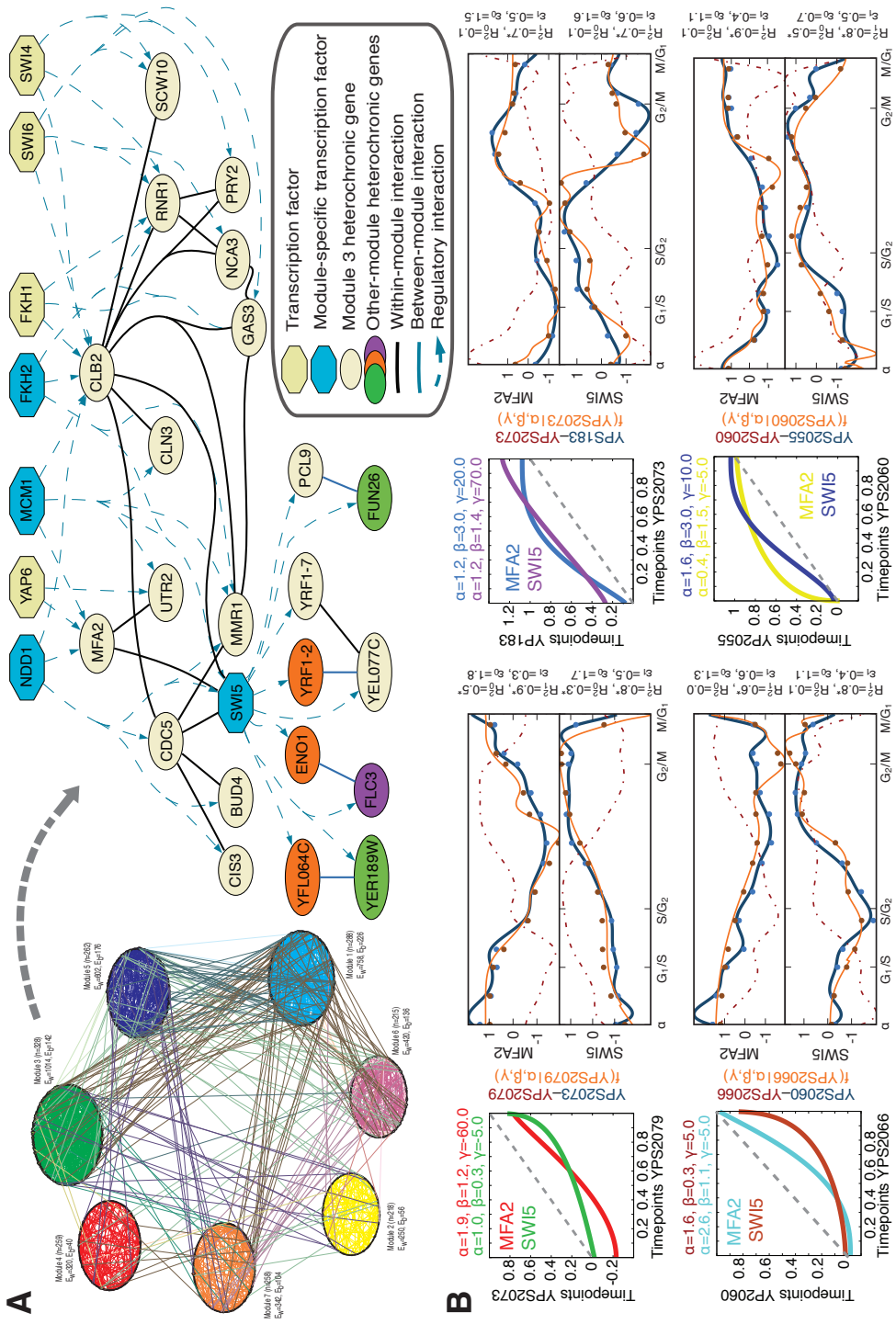


Figure 3.30: The modular architecture of genome-wide timing control

(A, left) Network of significant heterochronic interactions between 1828 core timing module genes, grouped by module. Figure 3.29 shows this graph at full resolution. **(A, right)** Heterochronic interaction network of module 3 (black lines); only the subset of genes within 2 degrees of gene Swi5 that share TFs is shown (dashed blue arrows). Interactions are defined by strongly correlated changes in expression timing. Swi5 itself encodes a module-specific TF. Blue nodes indicate significant association of a TF with a module. **(B)** Novel interaction between Swi5 and Mfa2, which co-cluster in 23/45 comparisons ($P = 6.8 \times 10^{-6}$); four are shown. Timing maps (columns 1,3) illustrate timing pattern changes between strains for each gene, given parameters (α, β, γ) and Beta CDF: $t' = (Beta(\alpha, \beta) + \gamma) \bmod 1$. Gray dashed lines indicate no change. Trajectory plots for each gene (columns 2,4) show the time transformation of CDC-expression from one strain (dashed red line) to another (orange line). Blue lines show a gene's CDC-expression in the respective target strain. Transformation order is reversible, since timepoint maps are invertible. R^2 and RMSE fit statistics are shown. *indicates significance ($P < 0.05$).

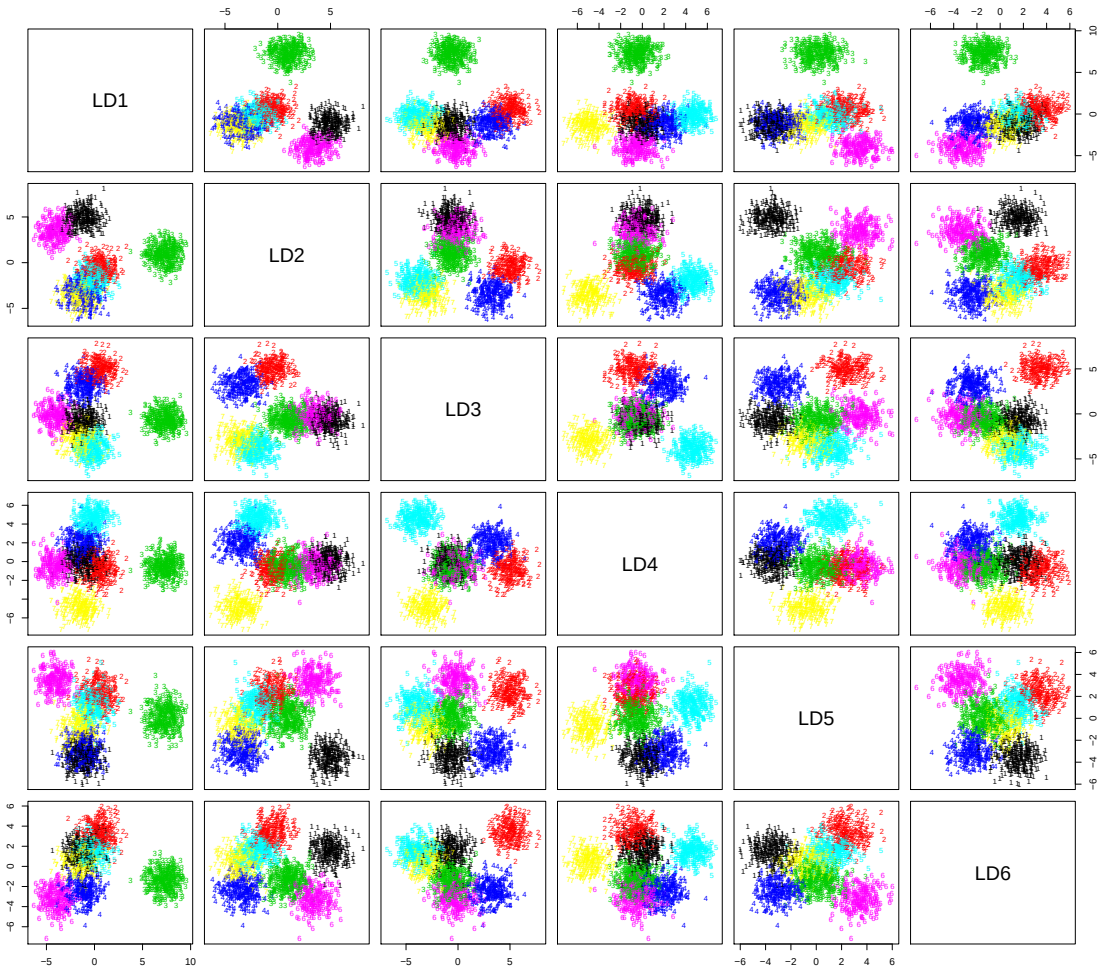


Figure 3.31: Linear discriminant visualization of 7 timing pattern clusters

A series of LDA plots are shown, illustrating 2D projections of the clustering of 1828 genes closely associated with each cluster. Clustering was obtained by k-means using 100 random starts. k was determined by gap statistic analysis (Figure 3.26A). Although difficult to visualize all together from a single low-dimensional projection, clusters are discriminated in 6 dimensions.

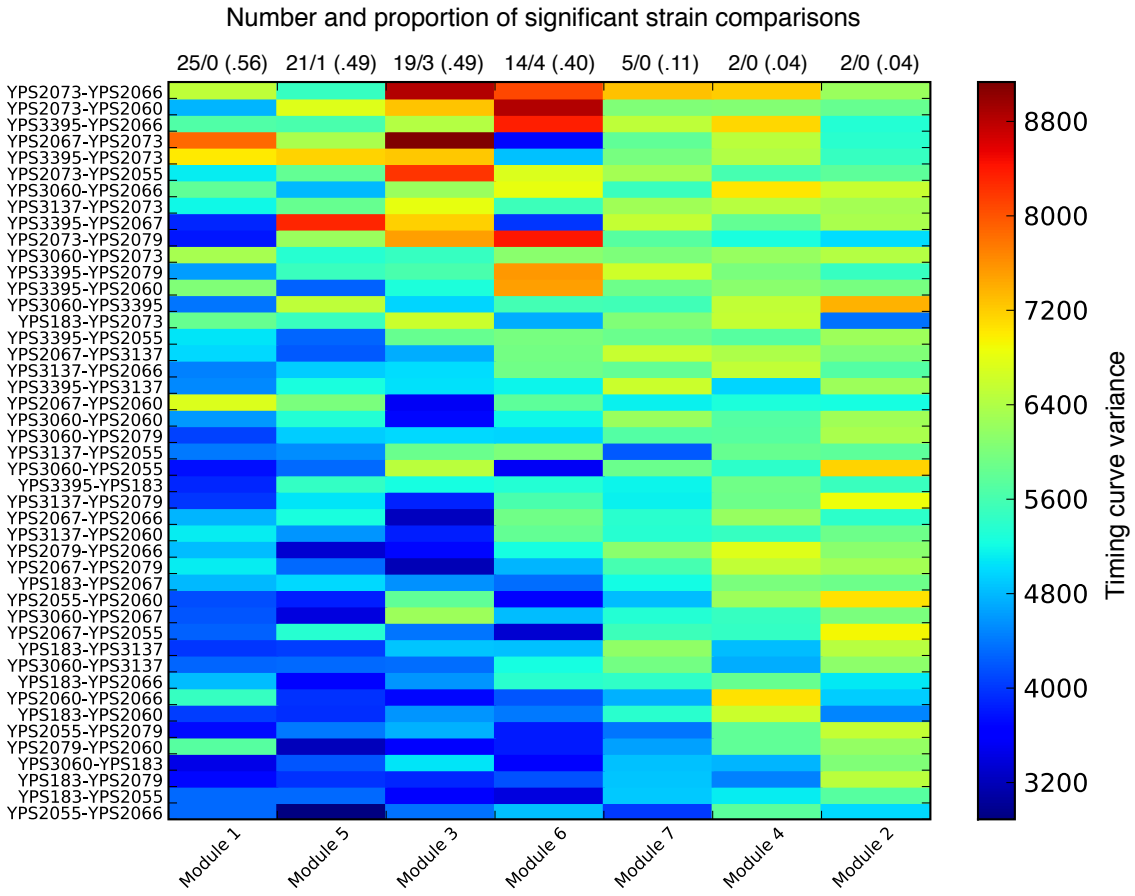


Figure 3.32: Heat map of modular expression timeline variance

Each cell illustrates the variance in timing patterns for the genes in each of 7 timing modules for one of 45 strain comparisons. Modules and comparisons are sorted by this variance. The number (and total proportion) of significantly low and high strain comparisons per module is shown (FDR < 0.001). On average 31% of comparisons are significant.

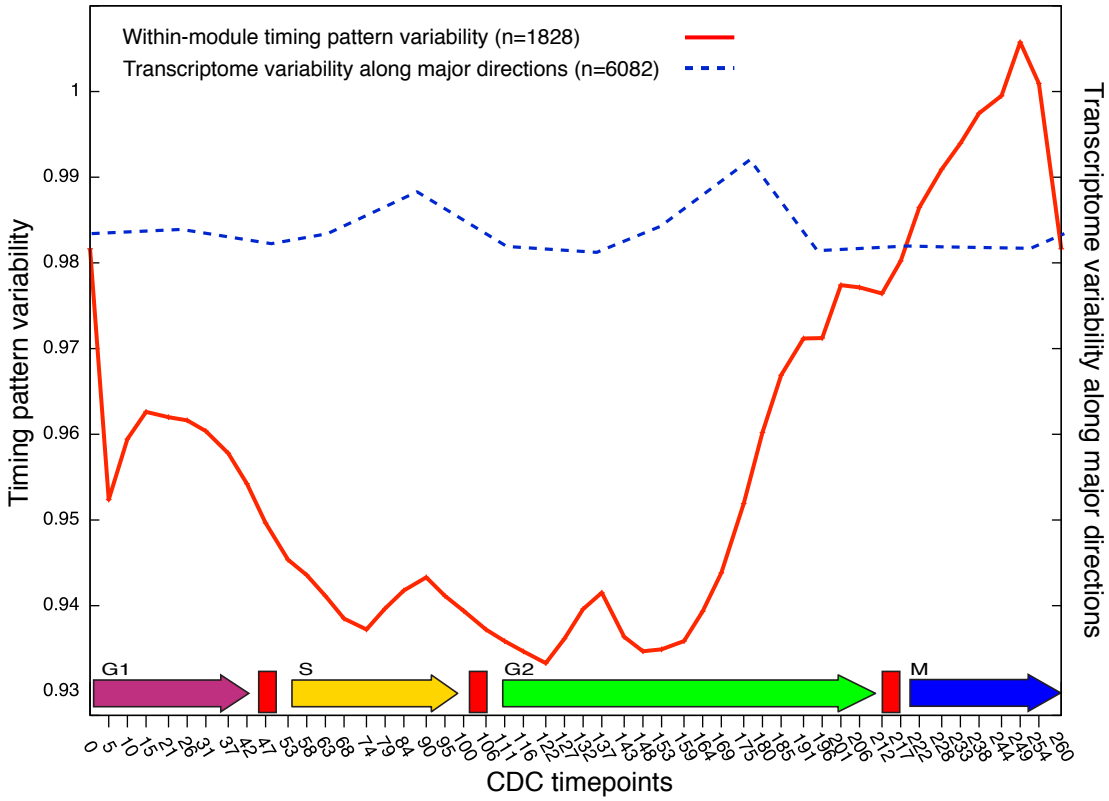


Figure 3.33: Gene expression timeline variability across the CDC for 1828 timing module genes

At each timepoint, variance across the timing patterns of within-module genes was computed. Variance values were then scaled by the expected variance at that timepoint, computed as the median of a distribution of 1000 random timing pattern variances evaluated at that timepoint. Estimated CDC-phases and checkpoints are shown at the bottom (taken from Figure 3.14). Dashed blue line shows transcriptome variability along the major CDC-directions (taken from Figure 3.14).

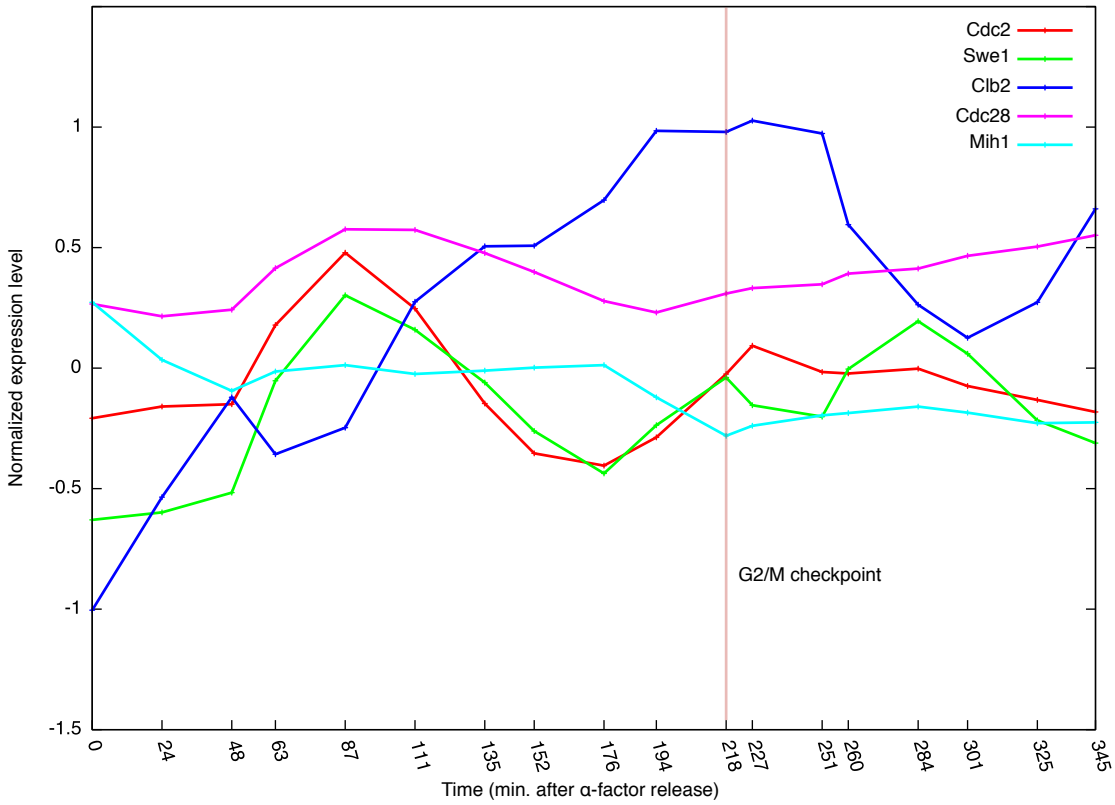


Figure 3.34: CDC-expression trajectories of genes involved in the timing of the G_2/M-phase transition

Data correspond to normalized gene expression levels measured in the laboratory strain YPS183. At timepoint 218 min., expression of the Y19-kinase Swe1 (YJL187C, inhibitor of Clb2-Cdc28 activity) peaks in G_2 (*35, 36*). This timepoint is coincident with nearly maximal levels of the B-type cyclin Clb2 (YPR119W) and low expression of the Y19-phosphatase Mih1 (YMR036C), a combination which promotes G_2/M progression (*37, 38*). Cdc2 (YDL102W), a gene whose deletion arrests CDC progression at the G_2/M checkpoint, also shows nearly maximal expression at this time.

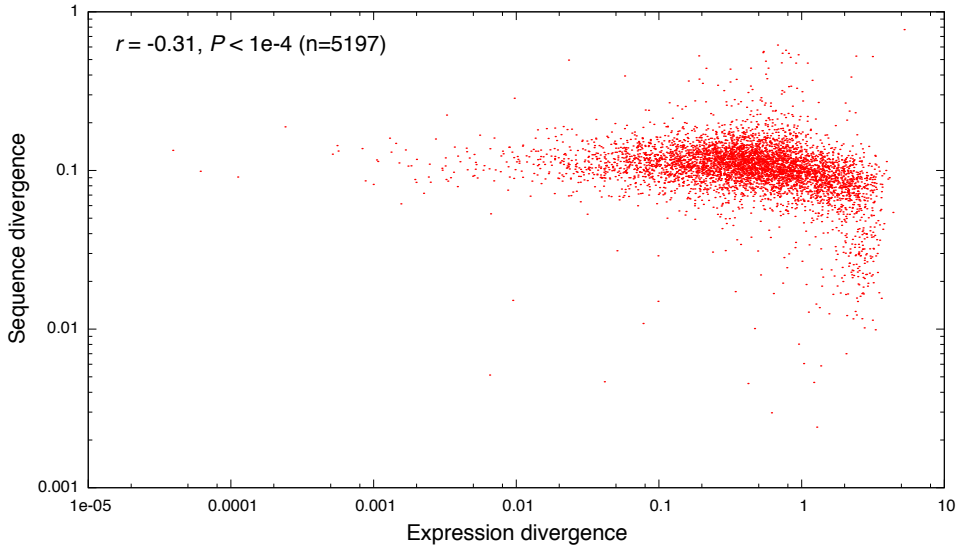


Figure 3.35: Comparison of sequence divergence and expression divergence between *S. cerevisiae* and *S. paradoxus*

Expression divergence was defined as the absolute difference in the time-averaged expression levels of each gene between species. Sequence divergence was defined as the average pairwise difference of each gene's coding sequence between species, normalized by coding sequence length. Sequence divergence was determined using global pairwise alignments of each gene. Coding sequences correspond to the laboratory strain *S. cerevisiae* strain S288c and *S. paradoxus* strain NRRL Y-17217. Expression data correspond to YPS183 (S288c derivative) and YPS3395 (woodland *S. paradoxus* strain). The lack of positive correlation and low explained variance (9%) indicate that divergence of *S. paradoxus* transcribed sequences will not overestimate expression variation between species.

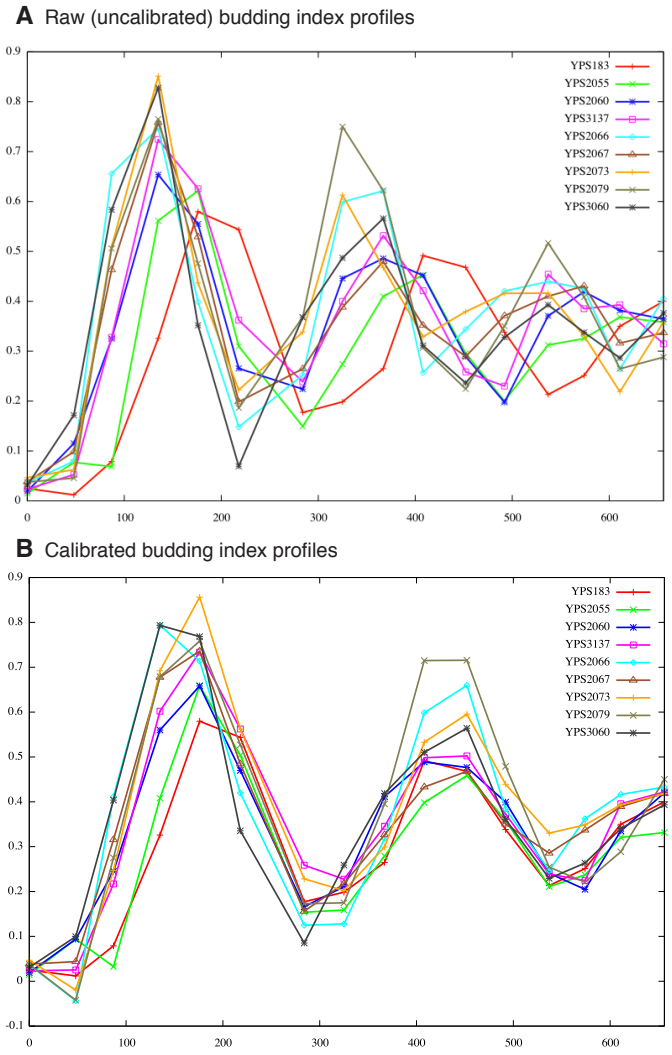


Figure 3.36: Budding index profiles for natural *S. cerevisiae* strains.

(A) Raw, uncalibrated budding index measurements across 2 cell-cycles for each strain. **(B)** Budding index profiles calibrated to the YPS183 curve using a linear scaling procedure (see Supplementary Methods). The budding index value at each timepoint is the average proportion of cells having two distinct nuclei under DAPI-filtered fluorescence light, assuming that under normal light, a single identifiable bud was found attached to a mother cell. At least 200 cells were counted on each slide. Measurements were taken following α-factor synchronization of each strain according to the protocol described in Materials and Methods.

Table 3.1: Natural and laboratory budding yeast isolates used in this study

ID	Strain ID	Species	Origin
1	YPS2055	*S. cerevisiae*	Tyler Arboretum, PA
2	YPS2060	*S. cerevisiae*	Mettlers Woods, NJ
3	YPS2066	*S. cerevisiae*	Mettlers Woods, NJ
4	YPS2067	*S. cerevisiae*	Tyler Arboretum, PA
5	YPS2073	*S. cerevisiae*	Mettlers Woods, NJ
6	YPS2079	*S. cerevisiae*	Westtown School Woods, PA
7	YPS3060	*S. cerevisiae*	Jenkins Woods, PA
8	YPS3137	*S. cerevisiae*	Jenkins Woods, PA
9	YPS183	*S. cerevisiae*	Laboratory (BY4741 derivative)
10	YPS3395	*S. paradoxus*	Jenkins Woods, PA

Isolates are haploid MATa, with the HO endonuclease locus replaced with a Kanamycin resistance cassette. YPS183 is also leu2Δ.

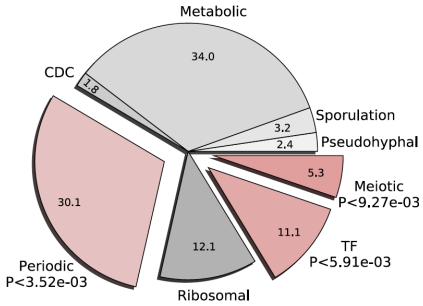
Table 3.2: Ranking of GO terms by proportion of associated genes evolving under stabilizing selection

Rank	GO-Slim term	Q	N	Rank	GO-Slim term	Q	N
1	helicase activity*	0.762	71	45	sporulation	0.97	95
2	extracellular region*	0.855	18	46	cellular respiration	0.97	76
3	cell wall*	0.912	82	47	cofactor metabolic process	0.97	152
4	cellular component*	0.915	748	48	nuclear organization and biogenesis	0.97	45
5	pseudohyphal growth*	0.934	50	49	RNA metabolic process	0.971	730
6	biological process	0.939	1162	50	mitochondrion	0.971	874
7	amino acid and derivative metabolic process	0.941	185	51	transferase activity	0.971	509
8	meiosis	0.941	94	52	cellular bud	0.971	122
9	plasma membrane	0.942	213	53	mitochondrial envelope	0.972	236
10	transporter activity	0.942	315	54	RNA binding	0.973	177
11	heterocycle metabolic process	0.943	73	55	site of polarized growth	0.973	123
12	microtubule organizing center	0.943	41	56	carbohydrate metabolic process	0.973	188
13	phosphoprotein phosphatase activity	0.95	40	57	endoplasmic reticulum	0.974	311
14	conjugation	0.953	89	58	cytoplasm	0.975	2294
15	molecular function	0.954	1718	59	cytoskeleton organization and biogenesis	0.975	173
16	signal transducer activity	0.954	40	60	structural molecule activity	0.975	285
17	cell wall org., biogen.	0.956	164	61	anatomical structure morphogenesis	0.975	170
18	cytokinesis	0.957	87	62	protein catabolic process	0.976	142
19	DNA metabolic process	0.957	520	63	protein folding	0.976	63
20	transcription regulator activity	0.958	238	64	cytoplasmic membrane-bound vesicle	0.976	82
21	lipid binding	0.958	38	65	peptidase activity	0.977	88
22	chromosome	0.958	175	66	electron transport	0.978	28
23	cellular homeostasis	0.958	116	67	nucleolus	0.979	181
24	cell cycle	0.959	306	68	transposition	0.979	8
25	vacuole	0.959	173	69	protein modification process	0.979	408
26	nucleotidyltransferase activity	0.959	52	70	endomembrane system	0.979	268
27	response to chemical stimulus	0.96	321	71	vitamin metabolic process	0.98	83
28	DNA binding	0.96	185	72	membrane organization and biogenesis	0.981	151
29	transport	0.961	764	73	Golgi apparatus	0.981	173
30	isomerase activity	0.961	51	74	translation regulator activity	0.981	47
31	response to stress	0.961	383	75	cell cortex	0.981	86
32	cytoskeleton	0.962	152	76	generation of precursor metabolites and energy	0.981	147
33	hydrolase activity	0.963	605	77	lipid metabolic process	0.981	204
34	nucleus	0.964	1350	78	ribosome biogenesis and assembly	0.983	289
35	ligase activity	0.964	129	79	translation	0.983	329
36	oxidoreductase activity	0.965	243	80	protein kinase activity	0.983	100
37	other	0.965	831	81	lyase activity	0.984	70
38	enzyme regulator activity	0.965	149	82	motor activity	0.984	14
39	membrane fraction	0.967	99	83	aromatic compound metabolic process	0.985	62
40	organelle organization and biogenesis	0.967	996	84	ribosome	0.985	301
41	transcription	0.968	426	85	vesicle-mediated transport	0.987	264
42	signal transduction	0.969	166	86	cell budding	0.989	68
43	membrane	0.969	770	87	peroxisome	0.994	47
44	protein binding	0.969	391	88	triplet codon-amino acid adaptor activity	.	0

88 GO-Slim terms are included, each of which is associated with a time-averaged proportion of genes of that term under stabilizing selection $\bar{Q}$, the number of associated genes N, and the rank of $\bar{Q}$, sorted from least to greatest proportion. This table corresponds to Figure 3.5. *indicates relative significance corresponding to the lowest 5% out of the distribution of 88 proportions.

Table 3.3: Genes with neutral or partly neutral expression trajectories for each life-cycle related term

Life-cycle term	N	Neutral genes
Periodic	35	SEO1, YBL112C, YCR018C-A, SIT1, MET6, DSE1, YFL064C, YFL065C, YFL066C, ECO1, MET10, YHL049C, YHI9, DSE2, YHR218W, YHR219W, YIL167W, YIL168W, YIL177C, YJL225C, MET3, ECM17, PIR1, YLL066C, YLL067C, PNP1, FRE1, CTS1, YLR462W, YRF1-5, RNH201, YNL165W, EGT2, YOR314W, YPR204W
Metabolic	14	IDP1, LYS21, EKI1, KGD2, MET6, ECO1, MET3, TIS11, MAG2, YRF1-5, TRM12, RNH201, NRK1, PUS4
Ribosomal	4	YDR266C, VMR1, NOP2, MOT1
CDC	3	FIG2, TIS11, CLB4
TF	2	ROX3, SUT2
Meiotic	1	FPR3
Pseudohyphal	1	MEP2
Sporulation	0	
		Partly neutral genes
Metabolic	129	PTA1, SCT1, SAS3, ORC2, RIB7, TAF5, HPC2, PYC2, DPB3, SNF5, ILV6, GLK1, RAD18, KNH1, NSE4, GLT1, TFP1, TRP1, LYS14, IPT1, FOB1, HPR1, HST4, SNU56, MET32, YDR287W, SUR2, IPK1, HIM1, SLD5, PCM1, PRP22, FIR1, HOM3, PRS2, IOC3, HXK1, PUF4, PYC1, CEG1, ARO8, SPT16, CAX4, ERG1, GCN5, SNF6, YNG2, HIS6, SER33, RPI1, KGD1, REV7, MET28, RNR2, RTT101, BNA3, RPE1, INO1, RFA3, ELO1, PRP21, HYS2, LSM8, TAH11, RAD7, CPA2, SOR1, CCE1, HCS1, ABF1, SRY1, PRP16, CDC45, MSL5, CKI1, SLX4, PUS5, ATG26, BNA5, TOP3, ACO1, SFH1, RSC2, SSQ1, IKI3, SWC7, RAD10, RNA14, RAD14, CEF1, YKU70, PUB1, PMS1, PHO23, LEU4, YAF9, PSD1, MET2, ARE2, POP2, OPI10, PSH1, ARG1, MSH2, EXO1, ALG8, TGL5, ELG1, LCB4, MRM1, SNF2, SPS4, PRO2, MIP1, MSC6, VTS1, RAD1, MET12, ERG10, YPL206C, LEA1, GAL4, CDC54, BRR1, THI22, ORC4, MET16, PRP4, SMX3
Periodic	114	KIN3, SAS3, SKT5, YBL111C, DSF2, TIP1, YBR089W, PHO89, AGP1, FUS1, GLK1, KAR4, HMLALPHA1, MATALPHA1, CPR4, PRM7, NSE4, YDL163W, MRH1, YDR053W, RVB1, HST4, SUR2, HXT7, NPL3, YHP1, MFA1, PLM2, GTT3, YEL075C, YEL076C, YEL076CA, SPC25, ISC1, FIR1, FTR1, YER189W, OCH1, YBP2, EMP24, SPT16, MUP1, DAM1, BUB1, BNS1, YHB1, SCW4, ARN1, SER33, PRM5, RPI1, MET28, BNA3, PRY1, CIS3, HSP150, RFA3, ELO1, SAG1, HYS2, SPC42, CWP1, YKL177W, STE3, UTH1, PXL1, CDC45, SLX4, BNA5, TOP3, YLR302C, YLR326W, YLR464W, SUR7, TEM1, PHO84, CTF18, NDE1, OCA2, PMS1, PSD1, APC1, YNL176C, YNL276C, YPT11, PFA3, MDJ2, DSE4, ARG1, YOL070C, EXO1, ELG1, MSB1, SLK19, WTM2, WTM1, YOR235W, HNT3, VPH1, MUM3, SPS4, YOR314W-A, PRO2, YPL025C, LEE1, HHO1, RDS2, YPL158C, BBP1, OPY2, YPR157W, MET16, YPR202W, YPR203W
Ribosomal	46	POP8, POP4, KRR1, RPP1A, NRP1, SED1, RRP8, EFT2, SNM1, GCN20, MRH4, SRM1, NSA1, RPS26A, HGH1, YGR198W, YGR251W, OTU2, PIH1, IPI1, RPF1, YIL096C, YIR003W, ALB1, YJL213W, RRP14, TRM2, TIF1, DRS1, EMG1, RPL31B, RPL6A, YMR114C, BCH1, RNT1, NIP1, KRE33, SKP2, ESF2, YOL070C, MDY2, EFT1, RPL33B, CAM1, TEF1, FHL1
TF	42	GAL1, TAF5, HPC2, SNF5, KAR4, HMLALPHA1, MATALPHA1, LYS14, YAP6, MHR1, PLM2, BUR6, HSF1, SPT16, TFC4, GCN5, YAP3, SNF6, GAT4, RSF2, IFH1, BDF1, TAF11, SOK2, STB2, TAF7, FAP1, SIN4, ESF2, GAL11, TOA1, WTM2, WTM1, SNF2, TAF3, RDS2, CUP9, GAL4, HAA1, MED1, FHL1, NUT2
Meiotic	20	SPO7, KAR4, YHP1, MAM1, ZIP2, BNS1, RIM4, SPO13, SET3, TOP3, BDF1, RAD50, SPO21, GAC1, SLK19, WTM2, WTM1, SPS4, MSC6, DDC1
Sporulation	12	DIT1, SPS1, RSC1, RIM4, SMC3, CRR1, YNL194C, RIM21, MPC54, SPR1, MUM3, SMK1
Pseudohyphal	9	CDC24, BUD5, STE7, NRG1, BMH2, NPL3, CDC42, SOK2, SRV2
CDC	7	SED1, CDC20, SMC3, AGA1, MSH2, CDC54, FHL1



Genes with neutral or partly neutral expression trajectories correspond to Figure 3.2D. Pie chart illustrating enrichment of the 742 partly neutral genes within each life-cycle related term **(bottom)** (*cf.* Figure 3.2E).

Table 3.4: Estimates time-dependent transcriptome coexpression structure

Time	BI	CDC-phase	Factors (*S.c.*)	Factors (*S.c.* + *S.p.*)	Interspecies $\angle$	F ($\times 10^{-3}$)	MDS d(t,t-1)
0	0.05	G_1	4	3	46.5	2.26	N/A
24	0.02	G_1	$\geq$2	$>$9	85.0	2.33	44.5°
48	0.06	$X_{G_1/S}$	$\geq$2	$>$9	86.1	2.09	59.5°
63	0.11	S	$\geq$7	$>$9	53.8	2.25	138.5°
87	0.26	S	7	$>$9	63.5	3.09	21.8°
111	0.44	X_{S/G_2}	8	2	85.9	2.05	41.4°
135	0.61	G_2	$\geq$7	$>$9	86.3	1.96	64.1°
152	0.67	G_2	$\geq$2	6	76.4	2.38	140.4°
176	0.69	G_2	$\geq$3	$>$9	83.7	3.93	19.6°
194	0.62	G_2	$\geq$8	$>$9	73.0	1.98	96.4°
218	0.47	$X_{G_2/M}$	$\geq$2	$\geq$3	80.1	2.06	125.7°
227	0.41	M	9	$\geq$4	86.0	2.04	8.1°
251	0.29	M	9	$\geq$6	87.0	2.02	126.4°
260	0.25	X_{M/G_1}	2	2	82.7	2.26	22.7°
284	0.18	G_1	$\geq$2	9	78.5	1.90	102.7°
301	0.18	$X_{G_1/S}$	$\geq$2	3	88.4	1.78	126.8°
325	0.21	S	$\geq$4	$>$9	75.2	1.89	19.2°
345	0.26	S	$\geq$2	$\geq$3	85.4	2.63	16.4°

The BI (budding index) column indicates the estimated proportion of cells having completed S-phase at each sampled timepoint, averaged over the 9 *S. cerevisiae* strains. Approximate CDC-phases are taken from Figures 3.34 and 3.11. Rows closest to the major CDC checkpoints ($X_{G_1/S}$ and $X_{G_2/M}$) are highlighted. The Factor columns indicate the number of independent latent factors required to explain differences in expression due to (presumably) genetic variation at a given timepoint. These numbers were estimated using a factor analytic linear mixed model applied to expression data of each strain, grouped by timepoint. Estimates were obtained in two ways, using only 9 *S. cerevisiae* data, and using the combined set of *S. cerevisiae* and *S. paradoxus* data. F-values represent the ratio of natural to neutral expression variance projected onto the major CDC-direction at each timepoint, scaled by degrees of freedom (8 and 22, respectively). Interspecies angles characterize the relationship of the *S. paradoxus* displacement vector to the major *S. cerevisiae* CDC-direction of variation at each timepoint. MDS angles are reproduced from Figure 3.12B.

Table 3.5: Comparison of evolutionary covariance matrices across sequential CDC timepoints using the common principle components test

Time t (min.)	CPC(t, $t-1$)	χ^2	d.f.	p-value
24	Proport	92.657	44	0.0000
48	CPC(3)	39.767	21	0.0079
63	CPC(4)	43.344	26	0.0178
87	CPC(7)	50.770	35	0.0413
111	CPC(6)	52.046	33	0.0187
135	Proport	95.623	44	0.0000
152	CPC(7)	63.181	35	0.0024
176	CPC(1)	21.572	8	0.0058
194	CPC(7)	74.548	35	0.0001
218	CPC(3)	33.102	21	0.0451
227	CPC(4)	42.277	26	0.0230
251	CPC(1)	28.472	8	0.0004
260	CPC(2)	30.205	15	0.0112
284	CPC(6)	71.779	33	0.0001
301	CPC(6)	53.970	33	0.0121
325	CPC(5)	51.553	30	0.0085
345	CPC(5)	48.791	30	0.0165

Limited to 9 degrees of freedom (number of *S. cerevisiae* strains), expression data from each timepoint t was first projected onto that timepoint's top 9 CDC-directions $U^{1\text{–}9}(t)$; subsequently 9×9 covariance matrices were computed for each timepoint. The degrees of freedom restriction only permits consideration of less than 50% of the total evolutionary covariation at each timepoint. 'Time' denotes the t and $t-1$ timepoints involved in each comparison (for $t = 24$, $t-1 = 0$). 'CPC(t, $t-1$)' shows the best model for each common principle components test (CPC) (the smallest one yielding a significant p-value < 0.05, see Materials and Methods). χ^2 shows the likelihood ratio test statistic. Respective degrees of freedom (d.f.) and Chi-square test p-value are also shown.

Table 3.6: Gene enrichments and top 5 genes correlating with major and rank-2 eigengenes throughout the CDC

Rank-11

Time	Significant GO term(s)	Gene (Pearson's r)
0	Ribosomal	AAD4 (0.95), ICY1 (0.94), AAD14 (0.94), PPT1 (0.92), VMR1 (0.92)
24	*sans*-TATA genes	SMF3 (0.98), SKY1 (0.95), AFG1 (0.93), TUL1 (0.92), YHR212C (0.91)
48	N/A	VID21 (0.97), HSL7 (0.96), NST1 (0.96), DNF1 (0.95), MAS6 (0.94)
63	CDC	NAP1 (0.97), SRO9 (0.92), YBR261C (0.90), YCR076C (0.90), FKH1 (0.88)
87	TATA genes	YBR139W (0.95), YJL028W (0.94), YNL305C (0.93), YML048W-A (0.92), LYS14 (0.91)
111	Periodic	GPI2 (0.96), APL1 (0.95), YCL026C-A (0.94), ARO8 (0.92), FUN26 (0.88)
135	N/A	PTH1 (0.89), REG1 (0.88), YIL161W (0.88), MRPS17 (0.88), MRPL20 (0.87)
152	N/A	SOL1 (0.96), YFR042W (0.94), RBK1 (0.91), VPS55 (0.91), YDL133W (0.91)
176	N/A	BUD32 (0.91), ICY2 (0.90), RTF1 (0.89), RPL23A (0.87), RPS0A (0.87)
194	TATA genes	OM45 (0.96), PYK2 (0.96), VPS17 (0.95), PRX1 (0.94), ARA1 (0.93)
218	*sans*-TATA genes	YHR212C (0.96), REV3 (0.95), YDR541C (0.95), SMF3 (0.95), YLR108C (0.95)
227	Periodic, Metabolic, CDC	YEL023C (0.95), PET54 (0.95), YDR186C (0.93), POL32 (0.92), DSE3 (0.89)
251	TATA genes	RAD7 (0.95), ROM1 (0.95), YOL087C (0.92), UBP15 (0.92), HST4 (0.91)
260	Ribosomal, *sans*-TATA genes	ACB1 (0.94), RSC4 (0.92), RPP1 (0.91), RPN2 (0.91), SER3 (0.90)
284	CDC	HHT2 (0.95), GET3 (0.93), TFA2 (0.92), GRX3 (0.90), SGT2 (0.89)
301	Metabolic, TATA genes, Periodic	ACO2 (0.96), ERG11 (0.96), LYS1 (0.93), UBP7 (0.92), TPN1 (0.91)
325	TATA genes, Metabolic	GCN5 (0.96), MET3 (0.95), YFL063W (0.92), PKC1 (0.92), YIL067C (0.92)
345	*sans*-TATA genes	NAM2 (0.96), PEX31 (0.95), MDR1 (0.92), YNR040W (0.92), CBP1 (0.91)

Rank-2

Time	Significant GO term(s)	Gene (Pearson's r)
0	N/A	DAL81 (0.96), PLB3 (0.92), STP22 (0.91), YBR014C (0.90), SPP381 (0.90)
24	N/A	TOR2 (0.95), YGR291C (0.93), YGL185C (0.92), ERG27 (0.92), WSS1 (0.91)
48	N/A	YBL095W (0.91), GIS4 (0.90), HAL5 (0.86), PAU7 (0.82), YLR064W (0.81)
63	Periodic, *sans*-TATA genes	MNN5 (0.92), YMR209C (0.91), PEX27 (0.91), TAF5 (0.88), YOL086W-A (0.88)
87	TATA genes	RAV2 (0.98), YKL151C (0.95), YMR31 (0.92), KTR4 (0.92), YPC1 (0.91)
111	Ribosomal	YJR115W (0.97), YEL067C (0.97), YPL108W (0.95), YBR300C (0.94), YPL197C (0.94)
135	CDC, Periodic, Meiotic	IFH1 (0.91), YNL162W-A (0.89), YDR239C (0.88), PMI40 (0.87), TOS2 (0.85)
152	Ribosomal	IPP1 (0.95), AFR1 (0.92), YRB1 (0.92), ACA1 (0.89), YER087C-A (0.89)
176	Ribosomal, TATA genes	YNR071C (0.92), MSB3 (0.91), APE2 (0.90), YJR115W (0.89), YOL086W-A (0.88)
194	N/A	PRD1 (0.94), PSO2 (0.91), MRPL50 (0.90), BIM1 (0.87), PUP1 (0.86)
218	N/A	PHO8 (0.96), MEK1 (0.96), TOR2 (0.96), TYR1 (0.94), YDR387C (0.93)
227	N/A	YGR043C (0.96), YLR063W (0.94), CNM67 (0.93), ALD3 (0.93), YPR158W (0.93)
251	N/A	SPC2 (0.93), APQ13 (0.93), IES2 (0.92), RRP7 (0.92), RPB9 (0.91)
260	Periodic, Metabolic, CDC	DCW1 (0.96), LST7 (0.93), CBF2 (0.93), GWT1 (0.92), VPS1 (0.92)
284	Ribosomal	YOR240W (0.94), TRS31 (0.92), PHO2 (0.91), RPS4A (0.90), YPL080C (0.90)
301	*sans* TATA genes, Pseudohyphal, TF	DAL81 (0.90), YFR016C (0.89), YER130C (0.89), ACT1 (0.88), YAR029W (0.88)
325	Ribosomal, TF, Meiotic, Sporulation	SPF1 (0.93), GCN1 (0.93), GAS1 (0.92), SPT5 (0.89), HSL7 (0.89)
345	N/A	TPM2 (0.96), AYR1 (0.93), RHO3 (0.92), SEC28 (0.92), HSP31 (0.92)

Enrichments were computed for the 5% of genes whose expression profiles across strains correlate best with the respective eigengene at a given timepoint (FDR < 0.1). The pool of terms includes 8 life-cycle related terms as well as the collection of genes harboring TATA regulatory motifs and genes lacking such motifs. The top 5 individual genes are reported for each timepoint with their respective Pearson correlation coefficients.

Table 3.7: Gene enrichment of 88 GO-Slim terms for the top 5% of genes correlating with major eigengenes throughout the CDC

Time	Significant GO term(s)
0	ribosome biogenesis and assembly, nucleolus, cellular homeostasis
24	endoplasmic reticulum, transport, transporter activity, lipid metabolic process, membrane, cytoplasmic membrane-bound vesicle, ligase activity, protein modification process, endomembrane system
48	N/A
63	chromosome, structural molecule activity, cytoskeleton organization and biogenesis, cytoskeleton, cell cycle, microtubule organizing center, cellular component, organelle organization and biogenesis
87	transporter activity, cellular homeostasis
111	transporter activity, plasma membrane, vacuole, cellular homeostasis, amino acid and derivative metabolic process, transport, signal transduction, vitamin metabolic process
135	structural molecule activity, mitochondrion, ribosome, translation
152	carbohydrate metabolic process, cytoplasm, mitochondrial envelope, protein catabolic process, generation of precursor metabolites and energy, peptidase activity, mitochondrion, membrane
176	N/A
194	mitochondrion, generation of precursor metabolites and energy, carbohydrate metabolic process, mitochondrial envelope, cofactor metabolic process
218	transporter activity, transport
227	helicase activity, DNA metabolic process, cell cycle, chromosome, response to stress, DNA binding, enzyme regulator activity, cellular bud
251	vacuole, lipid metabolic process, plasma membrane, phosphoprotein phosphatase activity, cytoplasm, cellular homeostasis
260	nucleolus, ribosome biogenesis and assembly, RNA metabolic process, RNA binding, organelle organization and biogenesis, nucleus, conjugation
284	protein catabolic process, peptidase activity, hydrolase activity, other, response to stress, response to chemical stimulus
301	amino acid and derivative metabolic process, lyase activity, aromatic compound metabolic process, lipid metabolic process, oxidoreductase activity, ligase activity
325	amino acid and derivative metabolic process, transporter activity, cellular homeostasis, oxidoreductase activity, plasma membrane, vacuole
345	mitochondrion

Significance was assessed at FDR < 0.05. This is a recapitulation of the enrichment analysis of Table 3.6, using more specific group labels.

Table 3.8: Ranking of gene groups by total number of genes across the CDC associated with a particular group

Top 5% of genes					
	Eigengenes, rank-1			Eigengenes, rank-2	
Rank	GO term	Genes	Rank	GO term	Genes
1	*sans*-TATA genes	4020	1	*sans*-TATA genes	3954
2	TATA genes	990	2	TATA genes	930
3	Metabolic	952	3	Metabolic	870
4	Periodic	698	4	Periodic	692
5	Ribosomal	303	5	Ribosomal	426
6	TF	222	6	TF	222
7	Meiotic	100	7	Meiotic	122
8	CDC	79	8	CDC	89
9	Sporulation	61	9	Sporulation	71
10	Pseudohyphal	61	10	Pseudohyphal	54

Top 5 genes					
	Eigengenes, rank-1			Eigengenes, rank-2	
Rank	GO term	Genes	Rank	GO term	Genes
1	*sans*-TATA genes	65	1	*sans*-TATA genes	65
2	TATA genes	20	2	TATA genes	17
3	Metabolic	20	3	Periodic	16
4	Ribosomal	8	4	Metabolic	9
5	Periodic	8	5	TF	7
6	TF	4	6	Ribosomal	4
7	CDC	4	7	CDC	2
8	Pseudohyphal	1	8	Meiotic	1
9	Sporulation	0	9	Sporulation	0
10	Meiotic	0	10	Pseudohyphal	0

Rankings are shown among the top 5% **(top)** or the top 5 genes correlating with major or rank-2 eigengenes **(bottom)**. Gene groups include 8 life-cycle related terms as well as TATA genes and *sans*-TATA genes.

Table 3.9: Statistics describing evolutionary divergence of the modular yeast coexpression structure

Diameter (%)	Sig. modules (%)	Overlap (% of diameter)	Excess %
25 (0.4)	1507.3 (24.8)	1.2 (4.8)	4.39
100 (1.6)	1645.7 (27.0)	10.6 (10.6)	8.96
500 (8.2)	3220.4 (52.9)	88.0 (17.6)	9.38
880 (14.5)	3389.2 (55.7)	215.6 (24.5)	10.03
1314 (21.6)	3625.3 (59.6)	408.6 (31.1)	9.49
2500 (41.1)	3972.1 (65.3)	1207.5 (48.3)	7.20

Statistics are reported for different module diameters and shown as both number of genes k and percentage of genome (with 6082 genes). A module is defined for every gene as the set of its k top correlating genes by Pearson correlation of temporal expression profiles. Sig. modules reports the number and percentage of significant gene modules ($P < 1/250$) averaged over all pairs of strains. Overlap reports the number of genes overlapping for a given module between a pair of strains, at the specified diameter k, averaged over all significant modules and all pairs of strains. This is also shown in parentheses as a percentage of diameter. 'Excess' shows the excess percentage of overlap compared to random expectation using binomial sampling. The excess percentage averaged over all k is 8.24%.

Table 3.10: Top 50 heterochronic genes, ranked by timing pattern distortion

Rank	Gene (Alias)	$\overline{R^2_{H_1}}$	$\overline{R^2_{H_0}}$	Sig. F-tests (Prop.)	Distortion
1	YGL040C (HEM2)	0.625	0.188	33 (0.733)	97.247
2	YNR054C (ESF2)	0.593	0.188	28 (0.622)	97.017
3	YBR261C (TAE1)	0.587	0.081	35 (0.778)	95.850
4	YOR308C (SNU66)	0.585	0.116	32 (0.711)	95.707
5	YOR011W (AUS1)	0.585	0.153	33 (0.733)	94.618
6	YKL025C (PAN3)	0.572	0.140	30 (0.667)	93.722
7	YLR015W (BRE2)	0.612	0.152	31 (0.689)	91.839
8	YPR107C (YTH1)	0.601	0.127	34 (0.756)	91.652
9	YBR206W	0.592	0.152	31 (0.689)	91.319
10	YDR288W (NSE3)	0.553	0.100	29 (0.644)	90.854
11	YDR205W (MSC2)	0.594	0.142	34 (0.756)	90.591
12	YOR257W (CDC31)	0.569	0.150	28 (0.622)	90.495
13	YHR034C (PIH1)	0.596	0.163	29 (0.644)	90.407
14	YGL002W (ERP6)	0.612	0.116	38 (0.844)	90.119
15	YFL023W (BUD27)	0.565	0.119	33 (0.733)	89.993
16	YHL030W (ECM29)	0.669	0.196	33 (0.733)	89.551
17	YLR063W	0.584	0.121	35 (0.778)	89.473
18	YLR158C (ASP3-3)	0.597	0.130	32 (0.711)	89.148
19	YDR384C (ATO3)	0.563	0.117	30 (0.667)	89.022
20	YDL194W (SNF3)	0.608	0.149	35 (0.778)	88.986
21	YDR505C (PSP1)	0.621	0.222	32 (0.711)	88.905
22	YIL165C	0.627	0.175	31 (0.689)	88.821
23	YKL032C (IXR1)	0.624	0.185	33 (0.733)	88.774
24	YLR464W	0.600	0.149	37 (0.822)	88.722
25	YCR091W (KIN82)	0.584	0.166	31 (0.689)	88.672
26	YKL175W (ZRT3)	0.573	0.148	29 (0.644)	88.417
27	YGR168C	0.610	0.122	34 (0.756)	88.355
28	YLR366W	0.572	0.156	29 (0.644)	88.349
29	YGL082W	0.569	0.117	33 (0.733)	88.146
30	YBR278W (DPB3)	0.587	0.161	33 (0.733)	88.100
31	YML009C-A	0.610	0.229	33 (0.733)	87.927
32	YJL020C (BBC1)	0.586	0.172	28 (0.622)	87.836
33	YHR018C (ARG4)	0.618	0.154	33 (0.733)	87.831
34	YJR012C	0.582	0.206	23 (0.511)	87.716
35	YPR126C	0.598	0.161	35 (0.778)	87.696
36	YNL008C (ASI3)	0.580	0.134	31 (0.689)	87.675
37	YDL006W (PTC1)	0.595	0.116	34 (0.756)	87.671
38	YHR137W (ARO9)	0.606	0.140	34 (0.756)	87.663
39	YBR123C (TFC1)	0.582	0.154	30 (0.667)	87.568
40	YNL001W (DOM34)	0.610	0.112	42 (0.933)	87.543
41	YLR282C	0.554	0.136	29 (0.644)	87.509
42	YIR033W (MGA2)	0.626	0.150	37 (0.822)	87.305
43	YMR242C (RPL20A)	0.599	0.151	31 (0.689)	87.201
44	YOL003C (PFA4)	0.615	0.154	33 (0.733)	87.201
45	YPR109W	0.571	0.103	36 (0.800)	87.129
46	YOR028C (CIN5)	0.582	0.165	29 (0.644)	87.101
47	YOR071C (NRT1)	0.622	0.151	36 (0.800)	87.035
48	YLR239C (LIP2)	0.618	0.140	37 (0.822)	86.746
49	YMR185W	0.615	0.197	34 (0.756)	86.709
50	YEL035C (UTR5)	0.594	0.140	36 (0.800)	86.648

$\overline{R^2_{H_1}}$ and $\overline{R^2_{H_0}}$ indicate the explained CDC-expression variation averaged over 45 strain comparisons, computed by time-domain or time-independent models, respectively. 'Sig. F-tests (prop.)' indicates the number (and proportion) of significant F-tests supporting the time-domain model, among strain comparisons. 'Distortion' indicates the RMSE of the optimal time-domain transformation curve against a line ($\alpha = 1, \beta = 1, \gamma = 0$), averaged over strain comparisons. The average distortion over all 6082 genes is 67.6 with a standard deviation of 8.7. The minimum distortion is 24.6 (GLN1). See Figure 3.23 for additional examples.

Table 3.11: Enrichment of functional ontology terms in timing modules, using the set of 1828 within-module genes

Module	Life-cycle term(s)	GO-slim term(s)
1 (n=288)	Ribosomal** (0.15)	cytoplasm** (0.56), translation** (0.13), ribosome** (0.13), structural molecule activity** (0.11), carbohydrate metabolic process** (0.08)
2 (n=218)	N/A	N/A
3 (n=328)	Periodic** (0.27)	cellular bud** (0.06), mitochondrion* (0.22), transporter activity* (0.1), site of polarized growth* (0.06), cell wall* (0.04), extracellular region* (0.02), electron transport* (0.02)
4 (n=259)	N/A	N/A
5 (n=262)	Ribosomal** (0.15), Metabolic+ (0.21)	ribosome biogenesis and assembly** (0.13), nucleolus** (0.11), nucleus+ (0.36), RNA metabolic process+ (0.23), oxidoreductase activity+ (0.08)
6 (n=215)	Ribosomal** (0.23)	ribosome biogenesis and assembly** (0.19), nucleolus** (0.11), RNA metabolic process+ (0.24), translation+ (0.11)
7 (n=258)	N/A	N/A

Life-cycle and GO-slim terms, along with their significance level, are listed for each timing module, using terms associated with the within-module genes of each module. Similar analysis of the set of between-module genes did not reveal any significantly enriched terms (FDR < 0.1). **indicates FDR < 0.01; *indicates FDR < 0.05; $^{+}$indicates FDR < 0.1.

Table 3.12: Modular expression timeline variability for 7 timing modules

Module	Rank	Within-species Low/high (prop.)	Between-species Low/high (prop.)	Total Low/high (prop.)
1	1	20/0 (0.56)	5/0 (0.56)	25/0 (0.56)
2	6	2/0 (0.06)	0/0 (0.00)	2/0 (0.04)
3	3	19/3 (0.61)	0/0 (0.00)	19/3 (0.49)
4	7	2/0 (0.06)	0/0 (0.00)	2/0 (0.04)
5	2	19/0 (0.53)	2/1 (0.33)	21/1 (0.49)
6	4	13/3 (0.44)	1/1 (0.22)	14/4 (0.40)
7	5	5/0 (0.14)	0/0 (0.00)	5/0 (0.11)

Variance in timing patterns among heterochronic genes was computed for each timing module and compared to a random variance distribution. Each random variance is the variance over a set of n random timing patterns, where n is the number of genes in a timing module. Random timing patterns were generated by choosing α, β, and γ values from the empirical distribution of estimated parameter values (see Figure 3.21). Significance was assessed across module for each of 45 strain comparisons using FDR < 0.001. The total numbers of significantly low and high comparisons are shown, along with the combined proportion of significant comparisons, for all 45 comparisons, the 36 within-species (*S. cerevisiae*) comparisons, and the 9 between-species comparisons.

Table 3.13: Heterochronic evolution in module-specific transcription factors

	Gene (Alias)	$\overline{R^2_{H_1}}$	$\overline{R^2_{H_0}}$	Sig. F-tests (Prop.)	Distortion
Module 1	YOR028C (CIN5*)	0.582	0.165	29 (0.644)	87.101
	YNL216W (RAP1**)	0.624	0.175	33 (0.733)	73.917
	YDR207C (UME6*)	0.532	0.153	27 (0.600)	63.198
	YKL062W (MSN4*)	0.706	0.242	36 (0.800)	54.906
Module 2	YER028C (MIG3*)	0.561	0.183	26 (0.578)	75.581
	YFL031W (HAC1*)	0.641	0.181	32 (0.711)	70.103
	YJR140C (HIR3*)	0.631	0.138	41 (0.911)	56.874
	YHR084W (STE12*)	0.577	0.142	30 (0.667)	55.373
	YMR019W (STB4**)	0.579	0.099	35 (0.778)	54.780
Module 3	YBL021C (HAP3*)	0.575	0.162	30 (0.667)	81.016
	YOR372C (NDD1**)	0.592	0.120	34 (0.756)	73.022
	YNL068C (FKH2***)	0.601	0.119	34 (0.756)	71.614
	YMR043W (MCM1*)	0.672	0.203	40 (0.889)	69.359
	YGL237C (HAP2*)	0.537	0.138	22 (0.489)	69.011
	YLR013W (GAT3*)	0.575	0.119	31 (0.689)	61.291
Module 4	YJR060W (CBF1*)	0.605	0.140	35 (0.778)	73.154
	YBL008W (HIR1**)	0.590	0.129	32 (0.711)	69.549
	YDL106C (PHO2*)	0.569	0.143	27 (0.600)	61.630
Module 5	YJL206C (YJL206C***)	0.571	0.111	34 (0.756)	75.654
	YMR075W (RCO1*)	0.616	0.128	36 (0.800)	50.028
Module 6	YDR009W (GAL3*)	0.605	0.121	35 (0.778)	71.873
Module 7	YGL071W (RCS1**)	0.581	0.115	33 (0.733)	73.417
	YOR162C (YRR1*)	0.605	0.137	33 (0.733)	72.379
	YDR146C (SWI5*)	0.764	0.289	41 (0.911)	63.564
Modules 1,6	YPR104C (FHL1***)	0.547	0.120	28 (0.622)	59.814
All modules	YJL056C (ZAP1*)	0.580	0.116	37 (0.822)	53.590

Chi-square tests were performed using 169 transcription factors (TFs) and each of the 7 timing modules with genes from the set of 1828 within-module genes. TFs that show significant enrichment either in particular modules compared to all others are shown; the last row shows a single TF (ZAP1) which is significantly associated with the pooled set of 1828 genes compared to non-heterochronic genes. ***indicates $P < 0.001$; **indicates $P < 0.01$; *indicates $P < 0.05$. There are 25 significant module-specific TFs, 1 of which associates with 2 modules (FHL1). In addition, 1 TF associates with the set of 1887 genes with patterns of heterochronic interaction between modules compared to the set of within-module genes (ZAP1). TF regulatory binding data were obtained from (*5*) using a cutoff of $P < 0.001$. $\overline{R^2_{H_1}}$ and $\overline{R^2_{H_0}}$ indicate the explained CDC-expression variation averaged over 45 strain comparisons, computed by the time-dependent heterochrony or time-independent models, respectively. The column Sig. F-tests (prop.) indicates the number (and proportion) of significant F-tests supporting the heterochrony model, among strain comparisons. Distortion indicates the RMSE of the optimal time transformation curve against a line ($\alpha = 1, \beta = 1, \gamma = 0$), averaged over strain comparisons. Genes are ranked by distortion for each category.

References

1. Rifkin, S. A., Kim, J., and White, K. P. *Nat Genet* **33**(2), 138–144 (2003).

2. Rifkin, S. A., Houle, D., Kim, J., and White, K. P. *Nature* **438**(7065), 220–223 (2005).

3. Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M., and Thomas, W. K. *Nat Genet* **37**(5), 544–548 (2005).

4. Lee, T. I., Rinaldi, N. J., Robert, F., et al. *Science* **298**(5594), 799–804 (2002).

5. Harbison, C. T., Gordon, D. B., Lee, T. I., et al. *Nature* **431**(7004), 99–104 (2004).

6. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. *Nature* **431**(7006), 308–312 (2004).

7. Stearns, S. C. and Magwene, P. *Am Nat* **161**(2), 171–180 (2003).

8. Beer, M. A. and Tavazoie, S. *Cell* **117**(2), 185–198 (2004).

9. Yuan, Y., Guo, L., Shen, L., and Liu, J. S. *PLoS Comput Biol* **3**(11), e243 (2007).

10. Prill, R. J., Iglesias, P. A., and Levchenko, A. *PLoS Biol* **3**(11), e343 (2005).

11. Alexander, R. P., Kim, P. M., Emonet, T., and Gerstein, M. B. *Sci Signal* **2**(81), pe44 (2009).

12. Olson, M. E. and Rosell, J. A. *Evolution Int J Org Evolution* **60**(4), 724–734 (2006).

13. Moss, E. G. *Curr Biol* **17**(11), R425–34 (2007).

14. Kim, J., Kerr, J. Q., and Min, G. S. *Proc Natl Acad Sci U S A* **97**(1), 212–216 (2000).

15. Somel, M., Franz, H., Yan, Z., Lorenc, A., Guo, S., Giger, T., Kelso, J., Nickel, B., Dannemann, M., Bahn, S., Webster, M. J., Weickert, C. S., Lachmann, M., Paabo, S., and Khaitovich, P. *Proc Natl Acad Sci U S A* **106**(14), 5743–5748 (2009).

16. Zeyl, C. and DeVisser, J. A. *Genetics* **157**(1), 53–61 (2001).

17. Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J., and Hartl, D. L. *Science* **317**(5834), 118–121 (2007).

18. Rifkin, S. A., Atteson, K., and Kim, J. *Funct Integr Genomics* **1**(3), 174–185 (2000).

19. Phillips, P. C. and Arnold, S. J. *Evolution* **53**(5), 1506–1515 October (1999).

20. Schluter, D. *Evolution* **50**(5), 1766–1774 (1996).

21. Pramila, T., Wu, W., Noble, W. S., and Breeden, L. http://www.fhcrc.org/science/labs/breeden/cellcycle/, (2008).

22. Gould, S. J. *Ontogeny and Phylogeny*. Harvard University Press, Cambridge, MA, (1977).

23. Alberch, P., Gould, S. J., Oster, G. F., and Wake, D. B. *Paleobiology* **5**(3), 296–317 (1979).

24. Bonner, J. T. *Size and Cycle: An Essay on the Structure of Biology*. Princeton University Press, (1965).

25. Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. *Proc Natl Acad Sci U S A* **101**(14), 4781–4786 (2004).

26. Shmulevich, I., Kauffman, S. A., and Aldana, M. *Proc Natl Acad Sci U S A* **102**(38), 13439–13444 (2005).

27. Wagner, G. P., Booth, G., and Bagheri-Chaichian, H. *Evolution* **51**(2), 329–347 (1996).

28. Stern, D. L. *Evolution* **54**(4), 1079–1091 August (2000).

29. Brauer, M. J., Huttenhower, C., Airoldi, E. M., et al. *Mol Biol Cell* **19**(1), 352–367 (2008).

30. Zhao, H., Butler, E., Rodgers, J., Spizzo, T., Duesterhoeft, S., and Eide, D. *J Biol Chem* **273**(44), 28713–28720 (1998).

31. Kuehne, H. A., Murphy, H. A., Francis, C. A., and Sniegowski, P. D. *Curr Biol* **17**(5), 407–411 (2007).

32. Spellman, P. T., Sherlock, G., Zhang, M. Q., et al. *Mol Biol Cell* **9**(12), 3273–3297 (1998).

33. Systat Software, Inc. *SigmaPlot Version 11*. Systat Software, Inc., San Jose, CA, (2007).

34. Vanoni, M., Vai, M., and Frascotti, G. *Cytometry* **5**(5), 530–533 (1984).

35. Booher, R. N., Deshaies, R. J., and Kirschner, M. W. *EMBO J* **12**(9), 3417–3426 (1993).

36. Sia, R. A., Bardes, E. S., and Lew, D. J. *EMBO J* **17**(22), 6678–6688 (1998).

37. Fitch, I., Dahmann, C., Surana, U., et al. *Mol Biol Cell* **3**(7), 805–818 (1992).

38. Sia, R. A., Herald, H. A., and Lew, D. J. *Mol Biol Cell* **7**(11), 1657–1666 (1996).

39. Lynch, M. and Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA, (1998).

40. Casella, G. and Berger, R. L. *Statistical Inference*. Duxbury, Pacific Grove, CA, second edition, (2002).

41. Fay, J. C. and Benavides, J. A. *PLoS Genet* **1**(1), 66–71 (2005).

42. Felsenstein, J. *Cladistics* **5**, 164–166 (1989).

43. Liti, G., Carter, D. M., Moses, A. M., et al. *Nature* **458**(7236), 337–341 (2009).

44. Simola, D. F. and Kim, J. In revision., (2008).

45. Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. *SAS System for Mixed Models*. SAS Institute, Cary, NC, (1996).

46. Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E., and Liu, J. S. *Nucleic Acids Res* **32**(2), 447–455 (2004).

47. Wolfram Research, Inc. *Mathematica Edition: Version 5.0*. Wolfram Research, Inc., Champaign, Il, (2003).

48. Tibshirani, R., Walther, G., and Hastie, T. *J. R. Statist. Soc. B.* **63**(2), 411–423 (2001).

Chapter 4

Conclusions and Future Directions

This dissertation offers a comprehensive survey of the evolution of genome-wide gene regulation among closely related yeast strains, including analyses of the evolution of the structural elements underlying gene expression—genes and their regulatory loci—and evolution of the continuous, time-dependent expression of these genes in the context of the mitotic cell-division cycle. These analyses suggest that negative selection is the dominant mode of evolution operating on both genetic changes (via purifying selection) and expression level changes (via stabilizing selection) genome-wide, consistent with the theory of stabilizing selection on developmental processes. In addition, elevated sequence divergence in promoter regions between species, the widespread divergence of expression dynamics within and between species, and the fact that heterochrony explains this divergence in expression dynamics for the majority of genes suggest that genome-wide gene regulation plays an important role in possible adaptive evolution among yeast strains and species. Thus although stabilizing selection appears to limit the overall evolution of expression levels, widespread heterochrony offers the potential for adaptive evolution via changes in expression dynamics. Moreover, changes in the scheduling and coordination of modular gene expression timelines, perhaps mediated pleiotropically by relatively few trans-regulatory factors, may be responsible for generating gene expression timing changes genome-wide. While these

results reveal novel evolutionary aspects of the yeast cell-division cycle, they also suggest a general architecture for temporal event control in biological systems comprised of a superposition of modular, dynamically-autonomous event schedules, each subject to different evolutionary forces. This modular dynamical architecture may facilitate the generation of combinatorially complex regulatory variation for evolutionary adaptation via changes in the scheduling and coordination of discrete event timelines, while buffering expression level changes in individual genes.

In chapter 2, I presented a genome-wide analysis of yeast evolutionary dynamics, using population genetic methodology to provide estimates of polymorphism and divergence based on sequence alignments of genic structural regions associated with the protein-coding genes of *S. cerevisiae* and *S. paradoxus*, using several dozen genomes. Despite the large number of genome sequences used to infer which evolutionary processes have influenced different genic structural regions, these conclusions may be tempered by the limitations of our analysis and of the data. Firstly, Tajima's D test, which was used to assess the effect of natural selection per-gene using polymorphic variation within species, is a statistically weak test because it relies on large sample sizes in order to detect rare alleles in a population. So despite having a few dozen genomes for each species, the somewhat limited number of sequences per-gene likely prevents the clear identification of modes of evolution for many genic structural regions. Secondly, the clonal and geographic population structures inherent to *S. cerevisiae* and *S. paradoxus* confound the detection of natural selection. Ideally, available genome sequences could be separated into smaller groups, within which random mating could be safely assumed. However, this would further reduce sample size, potentially leading to even weaker predictions. Until additional genomes are sequenced, it will be difficult to control for demographics properly and still obtain strong results. Thirdly, evolution of protein-coding sequences in particular deserves additional attention, not only

because they display the most sequence diversity as a structural class (since they encode the variety of molecular structures in a cell), but also because of their unique constraints with respect to the genetic code. Taking these limitations into consideration, I maintain that these data support the conclusion of genome-wide purifying selection. The yeast genomes analyzed are overall highly related within species, with 98.0% and 97.1% average genome identities within *S. cerevisiae* and *S. paradoxus*, respectively. This suggests that these strains likely hold in common the vast majority of their gene functions, and thus selection pressures as well. In addition, isolates from both species inter-breed with high efficiency (*1*), suggesting that few strain-specific epistatic effects have accumulated that might bias sequence comparisons (*e.g.* a locus with differential linkage relationships across strains may exhibit more or less extreme sequence variation). Both species are estimated to have large population sizes of 16 and 42 million organisms, enhancing the efficiency of natural selection. Lastly, although each species exhibits a unique population structure, our analysis found similar signatures of purifying selection in both species.

In chapter 3, I presented a comparative analysis of multi-genome time-series gene expression data for several closely related woodland yeast strains. While the entire process of gene expression is composed of many layers of regulation, we dealt exclusively with mRNA abundance, ignoring layers of translational regulation. Nevertheless, transcriptional regulation is an important and necessary component of gene expression, which, due to relatively fast mRNA decay rates of yeast genes (*2*), may be regarded as a limiting step in gene expression. For the analysis of these expression data, we adopted methods from quantitative genetics and multivariate statistics. Since the primary goal of the study was to identify time-dependent patterns of evolution, the majority of analyses incorporated time as a factor. However, deriving results from biological time-series data requires taking proper care in obtaining the measurements, due to the inherent tendency for organisms to vary. Since

yeast can be propagated asexually, it is easy to obtain large amounts of RNA expressed from a single genome. However, collecting suitable amounts of RNA for every gene at many closely spaced timepoints is much less straightforward. We chose to synchronize isogenic cultures of millions of haploid MATa cells using the α-factor mating pheromone, collecting genomic RNA samples as a time-series following the release of a culture from CDC arrest. Unfortunately, α-factor synchronization introduces an artificial and potentially significant environmental perturbation to the CDC, which may confound natural, physiological gene expression dynamics. However, visualization of the dominant genome-wide dynamics following α-factor perturbation suggests that its effect may be short-lived (*3*). A second consequence of synchronization may be the averaging of expression dynamics of individual cells. Viewed as a dynamical system, individual yeast cells could exhibit unique expression dynamics while progressing through the CDC. While dynamical variability has been observed among individual cells that are unsynchronized (*4*), the extent of overall dynamical variability in synchronized cultures is unclear. The fact that our data, in addition to other yeast time-series data (*5, 6, 7*), reveal clear periodic patterns in the expression of individual genes, suggests that variability in dynamics in a culture of synchronized cells may not be substantial.

Furthermore, in our association of regulatory factors with genes sharing similar expression dynamics, we utilized transcription factor–DNA binding data from Harbison *et al.* (*8*). Since these regulatory binding data reflect measurements made in a single laboratory yeast strain, the significant associations we identified might in some cases be spurious. Our results, however, suggest that the yeast regulatory structure could predominantly involve changes in the expression of relatively few regulatory factors rather than broad divergence of regulatory interactions. Moreover, changes in the expression of a transcription factor could result from changes in the binding sites in the promoter of the transcription factor,

nonsynonymous changes in its protein-coding sequence, or other upstream changes that change the timing of its expression. If changes in transcription factor expression typically result from minor modifications of these effects (*i.e.* effects which may alter but not eliminate the binding specificity of a transcription factor to DNA), the existing regulatory binding data should more or less accurately reflect the regulatory structure of many yeast strains. Otherwise, the existing regulatory binding data would only be accurate to a varying degree, depending on which and how many regulatory interactions had been altered. Thus, properly assessing which of the transcription factors identified in this study has the potential for causing downstream changes in expression dynamics at least requires sequencing these loci to determine whether genetic variation exists among strains. Our preliminary analysis of polymorphic variation in the proximal promoter regions of 8 transcription factor loci for these woodland strains (Table 1.1) does reveal elevated polymorphic variation in promoters compared to introns (Section 1.1.2). Since introns contain predominantly neutral sequence variation (Section 2.3.2), more polymorphic variation is expected among intronic sequences than among proximal promoters, which contain binding sequences for regulating gene expression and exhibit signatures of purifying selection (Section 2.3.4). Thus, finding more variation among promoter sequences than among introns is consistent with the role of genome-wide gene regulation in contributing to adaptive evolution of expression dynamics among closely related yeast strains.

4.1 Future Directions

Perhaps the most relevant questions raised by this dissertation pertain to the hypothesized modularity in the architecture of gene expression timing control, such as the relationship between this modular architecture and other modes of evolution as well as various environ-

mental perturbations. We found that this architecture consists of a set of 7 gene expression timing modules, each of which appears to execute a coherent event timeline. To arrive at this number of timing modules, we used changes in gene expression timing patterns across the CDC. Since our CDC-expression data set consists of a somewhat short and coarse-grained time-series ($\approx$19 min. resolution), these timing patterns were described by a minimal number of parameters. Thus our timing patterns may not fully capture all of the true changes in gene expression dynamics, in which case the true number of timing modules may be more than 7. In order to determine the number of timing modules more accurately, time-series gene expression data must be obtained at a much higher temporal resolution, so that a more capable time-domain transformation model might be employed to estimate a richer set of timing patterns.

In addition, we would like to verify independently the existence of this modular timing control architecture. One alternative would be to introduce global environmental perturbations, such as temperature change, to isogenic cell cultures and then to monitor gene expression dynamics genome-wide. Comparison of changes in expression dynamics across perturbations for a single genome would also be expected to reveal a modular timing control architecture. If the architecture of genome-wide gene regulation requires modularity of timing control for CDC progression, then changes in expression dynamics following these perturbations should reflect this modularity. Moreover, comparison of the changes in expression dynamics due to environmental change across different genomes might reveal evolutionary structural changes in the timing control architecture as well as corresponding genome-specific responses to the environmental perturbations, that is evolutionary changes in norms of reaction genome-wide.

The association of particular transcription factors with each timing module suggests that changes in the expression of one of these factors may pleiotropically affect downstream

expression dynamics of genes belonging to its timing module. One way to validate whether these TFs do specifically control expression timing in each module would be to perform under and over-expression studies of these TFs and monitor subsequent gene expression dynamics genome-wide. If changes in a TF's expression generally produce pleiotropic effects, then expression dynamics should change for a large number of genes. If these effects are segregated within a timing module, then only the expression dynamics of those genes within one timing module or whose expression is determined in part by that module should change.

Finally, comparison of changes in gene expression timing patterns (timelines) among the genes in each timing module revealed patterns of low variability. That is, timing modules appear to execute gene expression timelines in a coherent manner. Signatures of coherent expression dynamics suggest that the natural evolutionary variability in the timelines of individual genes belonging to a timing module may be limited by stabilizing selection. Moreover, it is the entire expression timeline that is limited across the CDC for many genes, indicating stabilization of a coherent developmental process. This suggests that the form of stabilizing selection involved may lead to canalization, the consequence of which would be a limited capacity to exhibit timeline variability within timing modules. A preliminary experiment would involve confirming whether the variance in timing modules is in fact evolutionarily limited compared to expectation from mutation–drift. To test this, one could obtain time-series expression data for several mutation accumulation lines and compute the expected timeline variation for each observed timing module. Observing less variation than expected would demonstrate that stabilizing selection operates on timing modules. To test explicitly for canalization, one also needs to show that the expected timeline variation in each module is also lower than what is possible barring any constraints on the production of timeline variation. One possible way to observe this is to obtain time-series expression

data for mutation accumulation lines at several different stages during their evolution from a common ancestor. If canalization does influence timing modules, there should be a progressive decay in its effect during the neutral evolution of each line, seen as an increase in timeline variance that correlates with the number of generations descendent from the common ancestor. This experiment is feasible and does not require time-series data across the entire CDC; rather expression data for several MA lines at a single, particular timepoint would suffice.

One predicted consequence of canalization is that it permits the genetic variation underlying a trait to be expressed to some degree, without affecting the trait's value. In this case there might be a sudden increase in mutational timeline variance after some number of generations, indicating that the mutational buffering capacity of a timing module has been overcome. In fact a test for environmental canalization might be performed similarly, using varying degrees of environmental perturbation to induce variation in timing modules (*9*).

4.2 Conclusion

A yeast cell's position in the cell-division cycle clearly influences observed patterns of evolution, in both magnitude and direction. The nature of this relationship involves dynamical changes in genome-wide gene regulation, which we have observed throughout the cell-division cycle. Although the time-dependent nature of the relationship between genome-wide gene regulation and evolution has been revealed, many questions remain before we understand more deeply the architecture of gene expression timing control defined by the yeast epigenotype.

References

1. Naumov, G. *Journal of Industrial Microbiology* **17**, 295–302 (1996).

2. Wang, Y., Liu, C. L., Storey, J. D., et al. *Proc Natl Acad Sci U S A* **99**(9), 5860–5865 (2002).

3. Rifkin, S. A. and Kim, J. *Bioinformatics* **18**(9), 1176–1183 (2002).

4. Bean, J. M., Siggia, E. D., and Cross, F. R. *Mol Cell* **21**(1), 3–14 (2006).

5. DeRisi, J. L., Iyer, V. R., and Brown, P. O. *Science* **278**(5338), 680–686 (1997).

6. Spellman, P. T., Sherlock, G., Zhang, M. Q., et al. *Mol Biol Cell* **9**(12), 3273–3297 (1998).

7. Pramila, T., Wu, W., Noble, W. S., and Breeden, L. http://www.fhcrc.org/science/labs/breeden/cellcycle/, (2008).

8. Harbison, C. T., Gordon, D. B., Lee, T. I., et al. *Nature* **431**(7004), 99–104 (2004).

9. Rutherford, S. L. and Lindquist, S. *Nature* **396**(6709), 336–342 (1998).

Works Cited

Aa, E., Townsend, J., Adams, R., Nielsena, K., and Taylor, J. Population structure and gene evolution in *Saccharomyces cerevisiae*. *FEMS Yeast Res* **6**, 702–715 (2006).

Alberch, P., Gould, S. J., Oster, G. F., and Wake, D. B. Size and shape in ontogeny and phylogeny. *Paleobiology* **5**(3), 296–317 (1979).

Alexander, R. P., Kim, P. M., Emonet, T., and Gerstein, M. B. Understanding modularity in molecular networks requires dynamics. *Sci Signal* **2**(81), pe44 (2009).

Baer, K. E. v. *Über Entwickelungsgeschichte der Thiere; Beobachtung und Reflexion*. Bei den Gebrüdern Bornträger, Königsberg, (1837).

Bean, J. M., Siggia, E. D., and Cross, F. R. Coherence and timing of cell cycle start examined at single-cell resolution. *Mol Cell* **21**(1), 3–14 (2006).

Beer, M. A. and Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**(2), 185–198 (2004).

Beer, M. A. and Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**(2), 185–198 (2004).

Boheler, K. Stem cell pluripotency: A cellular trait that depends on transcription factors, chromatin state and a checkpoint deficient cell cycle. *J Cell Physiol* **221**(1), 10–17 (2009).

Bonner, J. T. *Size and Cycle: An Essay on the Structure of Biology*. Princeton University Press, (1965).

Booher, R. N., Deshaies, R. J., and Kirschner, M. W. Properties of *Saccharomyces cerevisiae* Wee1 and its differential regulation of p34 CDC28 in response to G1 and G2 cyclins. *EMBO J* **12**(9), 3417–3426 (1993).

Brakefield, P. M. Evo-devo and constraints on selection. *Trends Ecol Evol* **21**(7), 362–368 (2006).

Brauer, M. J., Huttenhower, C., Airoldi, E. M., Rosenstein, R., Matese, J. C., Gresham, D., Boer, V. M., Troyanskaya, O. G., and Botstein, D. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Mol Biol Cell* **19**(1), 352–367 (2008).

Brigandt, I. Homology and heterochrony: the evolutionary embryologist Gavin Rylands de Beer (1899-1972). *J Exp Zoolog B Mol Dev Evol* **306**(4), 317–328 (2006).

Brown, R. and Danielli, J. F., editors. *Symposia Of The Society For Experimental Biology - Number VII Evolution*. Academic Press, New York, (1953).

Casella, G. and Berger, R. L. *Statistical Inference*. Duxbury, Pacific Grove, CA, second edition, (2002).

Chamary, J.-V. and Hurst, L. D. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: Evidence for selectively driven codon usage. *Mol Biol Evol* **21**(6), 1014–1023 (2004).

Chan, R. K. and Otte, C. A. Physiological characterization of *Saccharomyces cerevisiae* mutants supersensitive to G1 arrest by a factor and alpha factor pheromones. *Mol Cell Biol* **2**(1), 21–29 (1982).

Charlesworth, B., Morgan, M. T., and Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**(4), 1289–1303 (1993).

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**(5629), 71–76 July (2003).

Conant, G. C. and Wolfe, K. H. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* **179**(3), 1681–1692 (2008).

Cvijovic, M., Dalevi, D., Bilsland, E., Kemp, G. J. L., and Sunnerhagen, P. Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC Bioinformatics* **8**, 295 (2007).

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., and Steinmetz, L. M. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**(14), 5320–5325 (2006).

Davidson, E. H. and Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–800 10 February (2006).

De Beer, G. *Embryology and Evolution*. The Clarendon Press, Oxford, (1930).

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**(5710), 724–727 (2005).

Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M., and Thomas, W. K. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet* **37**(5), 544–548 (2005).

DeRisi, J. L., Iyer, V. R., and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**(5338), 680–686 (1997).

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.-M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.-M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.-F., Straub, M.-L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., and Souciet, J.-L. Genome evolution in yeasts. *Nature* **430**(6995), 35–44 (2004).

Saccharomyces Genome Database. http://yeastgenome.org, January (2008).

Farabaugh, P. J. Post-transcriptional regulation of transposition by Ty retrotransposons of *Saccharomyces cerevisiae*. *J Biol Chem* **270**(18), 10361–10364 May (1995).

Fay, J. C. and Benavides, J. A. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet* **1**(1), 66–71 (2005).

Fay, J. C. and Benavides, J. A. Hypervariable noncoding sequences in *Saccharomyces cerevisisae*. *Genetics* **170** August (2005).

Fay, J. C., McCullough, H. L., Sniegowski, P. D., and Eisen, M. B. Population genetic variation in gene expression is associated with phenotypic variation in saccharomyces cerevisiae. *Genome Biol* **5**(4), R26 (2004).

Fay, J. C. and Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics* **155**(3), 1405–1413 (2000).

Felsenstein, J. Phylip - phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).

Ferea, T. L., Botstein, D., Brown, P. O., and Rosenzweig, R. F. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc Natl Acad Sci U S A* **96**(17), 9721–9726 (1999).

Fingerman, E. G., Dombrowski, P. G., Francis, C. A., and Sniegowski, P. D. Distribution and sequence analysis of a novel Ty3-like element in natural *Saccharomyces paradoxus* isolates. *Yeast* **20**(9), 761–770 (2003).

Fischer, G., James, S. A., Roberts, I. N., Oliver, S. G., and Louis, E. J. Chromosomal evolution in *Saccharomyces*. *Nature* **405**(6785), 451–454 (2000).

Fitch, I., Dahmann, C., Surana, U., Amon, A., Nasmyth, K., Goetsch, L., Byers, B., and Futcher, B. Characterization of four B-type cyclin genes of the budding yeast *Saccharomyces cerevisiae*. *Mol Biol Cell* **3**(7), 805–818 (1992).

Frazer, K. A., Ballinger, D. G., Cox, D. R., et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164), 851–861 (2007).

Gadgil, M. and Bossert, W. H. Life historical consequences of natural selection. *Amer Nat* **104**(935), 1–24 (1970).

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**(12), 4241–4257 (2000).

Gerhart, J. and Kirschner, M. *Cells, Embryos, and Evolution: toward a Cellular and Developmental Understanding of Phenotypic Variation and Evolutionary Adaptability*. Blackwell Science, Malden, MA, (1997).

Gerstein, A. C., Chun, H.-J., Grant, A., and Otto, S. P. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet* **2**(9) (2006).

Giaever, G., Chu, A. M., Ni, L., et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**(6896), 387–391 (2002).

Gibson, G. and Wagner, G. Canalization in evolutionary genetics: a stabilizing theory? *Bioessays* **22**, 372–380 (2000).

Gilad, Y., Oshlack, A., and Rifkin, S. A. Natural selection on gene expression. *Trends Genet* **22**(8), 456–461 (2006).

Gilad, Y., Oshlack, A., Smyth, G. K., Speed, T. P., and White, K. P. Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* **440**(7081), 242–245 (2006).

Gilbert, S. F. *Developmental Biology*. Sinauer Associates, Inc., 6th edition, (2000).

Gilbert, S. F. Diachronic biology meets evo-devo: C. H. Waddington's approach to evolutionary developmental biology. *American Zoologist* **40**(5), 729–737 November (2000).

Gillespie, J. H. *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, second edition, (2004).

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. Life with 6000 genes. *Science* **274**(5287), 546, 563–7 (1996).

Goldschmidt, R. The theory of the gene. *The Scientific Monthly* **46**(3), 268–273 (1938).

Goldschmidt, R. *The Material Basis of Evolution*. Yale University Press, New Haven, CT, (1982).

Goll, J. and Uetz, P. The elusive yeast interactome. *Genome Biol* **7**(6), 223 (2006).

Gould, S. J. *Ontogeny and Phylogeny*. Harvard University Press, Cambridge, MA, (1977).

Gould, S. J. Is a new and general theory of evolution emerging? *Paleobiology* **6**(1), 119–130 (1980).

Haeckel, E. H. P. A. *Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie*. G. Reimer, Berlin, (1866).

Haerty, W. and Singh, R. S. Gene regulation divergence is a major contributor to the evolution of Dobzhansky–Muller incompatibilities between species of *Drosophila*. *Mol Biol Evol* **9**, 1707–1714 (2006).

Hall, B. K., Pearson, R. D., and Müller, G. *Environment, Development, and Evolution: toward a Synthesis*. MIT Press, Cambridge, MA, (2004).

Hamatani, T., Carter, M. G., Sharov, A. A., and Ko, M. S. H. Dynamics of global gene expression changes during mouse preimplantation development. *Dev Cell* **6**(1), 117–131 (2004).

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., and Young, R. A. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004), 99–104 (2004).

Hartwell, L. H., Culotti, J., Pringle, J. R., and Reid, B. J. Genetic control of the cell division cycle in yeast. *Science* **183**(4120), 46–51 (1974).

Hartwell, L. H. and Weinert, T. A. Checkpoints: controls that ensure the order of cell cycle events. *Science* **246**(4930), 629–634 (1989).

Herskowitz, I. Life cycle of the budding yeast *Saccharomyces cerevisiae*. *Microbiol Rev* **52**(4), 536–553 (1988).

Hooke, R. *Micrographia: or Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses with Observations and Inquiries thereupon*. John Martyn and James Allestry, November (1664).

Howe, K. J. and Ares, M. J. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc Natl Acad Sci U S A* **94**(23), 12467–12472 (1997).

Hudson, R. R., Kreitman, M., and Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 May (1987).

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. Functional discovery via a compendium of expression profiles. *Cell* **102**(1), 109–126 (2000).

Ingolia, N. and Murray, A. The ups and downs of modeling the cell cycle. *Curr Biol* **14** (2004).

Johnson, L. J., Koufopanou, V., Goddard, M. R., Hetherington, R., Schäfer, S. M., and Burt, A. Population genetics of the wild yeast *Saccharomyces paradoxus*. *Genetics* **166**, 43–52 January (2004).

Jorgensen, P., Rupes, I., Sharom, J. R., Schneper, L., Broach, J. R., and Tyers, M. A dynamic transcriptional network communicates growth potential to ribosome synthesis and critical cell size. *Genes Dev* **18**(20), 2491–2505 (2004).

Jorgensen, P. and Tyers, M. How cells coordinate growth and division. *Curr Biol* **13**, 1014–1027 December (2004).

Keightley, P. D., Lercher, M. J., and Eyre-Walker, A. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol* **3**(2) February (2005).

Kellis, M., Birren, B. W., and Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**(6983), 617–624 (2004).

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423** May (2003).

Kim, J., Kerr, J. Q., and Min, G. S. Molecular heterochrony in the early development of *Drosophila*. *Proc Natl Acad Sci U S A* **97**(1), 212–216 (2000).

Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**(5129), 624–626 (1968).

Kimura, M. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, (1983).

Klevecz, R. R., Li, C. M., Marcus, I., and Frankel, P. H. Collective behavior in gene regulation: the cell is an oscillator, the cell cycle a developmental process. *FEBS J* **275**(10), 2372–2384 (2008).

Koufopanou, V., Hughes, J., Bell, G., and Burt, A. The spatial scale of genetic differentiation in a model organism: the wild yeast *Saccharomyces paradoxus*. *Philos Trans R Soc Lond B Biol Sci* **361**(1475), 1941–1946 (2006).

Kuehne, H. A. *The Genetic Structure and Biogeography of Natural Saccharomyces Populations*. PhD thesis, University of Pennsylvania, (2005).

Kuehne, H. A., Murphy, H. A., Francis, C. A., and Sniegowski, P. D. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. *Curr Biol* **17**(5), 407–411 (2007).

Lande, R. Natural selection and random genetic drift in phenotypic evolution. *Evolution* **30**(2), 314–334 (1976).

Landry, C. R., Lemos, B., Rifkin, S. A., Dickinson, W. J., and Hartl, D. L. Genetic properties influencing the evolvability of gene expression. *Science* **317**(5834), 118–121 (2007).

Landry, C. R., Townsend, J. P., Hartl, D. L., and Cavalieri, D. Ecological and evolutionary genomics of *Saccharomyces cerevisiae*. *Mol Ecol* **15**(3), 575–591 (2006).

Laubichler, M. D. and Maienschein, J., editors. *From Embryology to Evo-Devo: A History of Developmental Evolution*. The MIT Press, Cambridge, MA, (2007).

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**(5594), 799–804 (2002).

Lewin, B. *Genes VII*. Oxford University Press, Oxford, (2000).

Li, F., Long, T., Lu, Y., Ouyang, Q., and Tang, C. The yeast cell-cycle network is robustly designed. *Proc Natl Acad Sci U S A* **101**(14), 4781–4786 (2004).

Li, W.-H. *Molecular Evolution*. Sinauer Associates, Sunderland, MA, (1997).

Libri, D., Lescure, A., and Rosbash, M. Splicing enhancement in the yeast *Rp51b* intron. *RNA* **6**(3), 352–368 (2000).

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., Davey, R. P., Roberts, I. N., Burt, A., Koufopanou, V., Tsai, I. J., Bergman, C. M., Bensasson, D., O'Kelly, M. J. T., van Oudenaarden, A., Barton, D. B. H., Bailes, E., Nguyen, A. N.,

Jones, M., Quail, M. A., Goodhead, I., Sims, S., Smith, F., Blomberg, A., Durbin, R., and Louis, E. J. Population genomics of domestic and wild yeasts. *Nature* **458**(7236), 337–341 (2009).

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. *SAS System for Mixed Models*. SAS Institute, Cary, NC, (1996).

Lopez, P. J. and Seraphin, B. YIDB: the yeast intron database. *Nucleic Acids Res* **28**(1), 85–86 (2000).

Lopez-Maury, L., Marguerat, S., and Bahler, J. Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet* **9**(8), 583–593 August (2008).

Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E., and Liu, J. S. Statistical resynchronization and bayesian detection of periodically expressed genes. *Nucleic Acids Res* **32**(2), 447–455 (2004).

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**(7006), 308–312 (2004).

Lynch, M. Phenotypic evolution by neutral mutation. *Evolution* **40**(5), 915–935 (1986).

Lynch, M. and Conery, J. S. The origins of genome complexity. *Science* **302**(5649), 1401–1404 (2003).

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C. R., Dopman, E. B., Dickinson, W. J., Okamoto, K., Kulkarni, S., Hartl, D. L., and Thomas, W. K. A genome-wide view

of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**(27), 9272–9277 (2008).

Lynch, M. and Walsh, B. *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA, (1998).

MacIsaac, K. D., Wang, T., Gordon, D. B., Gifford, D. K., Stormo, G. D., and Fraenkel, E. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 113 (2006).

Mayr, E. *Systematics and the Origin of Species*. Columbia University Press, New York, (1942).

Mortimer, R. K. Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res* **10**, 403–409 (2000).

Mortimer, R. K. and Hawthorne, D. C. Yeast genetics. *Annu. Rev. Microbiol.* **20**, 151–168 October (1966).

Mortimer, R. K., Romano, P., Suzzi, G., and Polsinelli, M. Genome renewal: a new phenomenon revealed from a genetic study of 43 strains of *Saccharomyces cerevisiae* derived from natural fermentation of grape musts. *Yeast* **10**(12), 1543–1552 (1994).

Moss, E. G. Heterochronic genes and the nature of developmental time. *Curr Biol* **17**(11), R425–34 (2007).

Murphy, H. A., Kuehne, H. A., Francis, C. A., and Sniegowski, P. D. Mate choice assays and mating propensity differences in natural yeast populations. *Biol Lett* **2**(4), 553–556 (2006).

Naumov, G. Genetic identification of biological species in the *Saccharomyces sensu stricto* complex. *Journal of Industrial Microbiology* **17**, 295–302 (1996).

Naumov, G. I., Naumova, E. S., and Sniegowski, P. D. *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* are associated with exudates of North American Oaks. *Can J Microbiol* **44**, 1045–1050 (1998).

Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**(5428), 96–98 (1973).

Oleksiak, M. F., Churchill, G. A., and Crawford, D. L. Variation in gene expression within and among natural populations. *Nat Genet* **32**(2), 261–266 (2002).

Olson, M. E. and Rosell, J. A. Using heterochrony to detect modularity in the evolution of stem diversity in the plant family *Moringaceae*. *Evolution Int J Org Evolution* **60**(4), 724–734 (2006).

Orlando, D. A., Lin, C. Y., Bernard, A., Wang, J. Y., Socolar, J. E. S., Iversen, E. S., Hartemink, A. J., and Haase, S. B. Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature* **453**(7197), 944–947 (2008).

Oyama, S., Griffiths, P. E., and Gray, R. D. *Cycles of Contingency: Developmental Systems and Evolution*. MIT Press, Cambridge, MA, (2001).

Phillips, P. C. and Arnold, S. J. Hierarchical comparison of genetic variance-covariance matrices. I. using the Flury hierarchy. *Evolution* **53**(5), 1506–1515 October (1999).

Pramila, T., Wu, W., Noble, W. S., and Breeden, L. Periodic genes of the yeast *Saccharomyces cerevisiae*: A combined analysis of five cell cycle data sets. http://www.fhcrc.org/science/labs/breeden/cellcycle/, (2008).

Prill, R. J., Iglesias, P. A., and Levchenko, A. Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol* **3**(11), e343 (2005).

Raff, R. A. *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. University of Chicago Press, Chicago, IL, (1996).

Raff, R. A. and Wray, G. A. Heterochrony: Developmental mechanisms and evolutionary results. *J Evol Biol* **2**, 409–434 (1989).

Replansky, T., Koufopanou, V., Greig, D., and Bell, G. *Saccharomyces sensu stricto* as a model system for evolution and ecology. *Trends Ecol Evol* **23**(9), 494–501 (2008).

Rice, S. H. *Encyclopedia of Evolution*, chapter Heterochrony, http://www.oxford–evolution.com/entry?entry=t169.e191. Oxford University Press, University of Pennsylvania, e-reference edition edition, (2005).

Rifkin, S. A., Atteson, K., and Kim, J. Constraint structure analysis of gene expression. *Funct Integr Genomics* **1**(3), 174–185 (2000).

Rifkin, S. A., Houle, D., Kim, J., and White, K. P. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* **438**(7065), 220–223 (2005).

Rifkin, S. A. and Kim, J. Geometry of gene expression dynamics. *Bioinformatics* **18**(9), 1176–1183 (2002).

Rifkin, S. A., Kim, J., and White, K. P. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat Genet* **33**(2), 138–144 (2003).

Ringo, J. *Fundamental Genetics*. Cambridge University Press, New York, (2004).

Rockman, M. V. and Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* **7**(11), 862–872 (2006).

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* **24**(3), 227–235 (2000).

Ruderfer, D. M., Pratt, S. C., Seidel, H. S., and Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet* **38**(9), 1077–1081 (2006).

Rutherford, S. L. and Lindquist, S. Hsp90 as a capacitor for morphological evolution. *Nature* **396**(6709), 336–342 (1998).

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. Positive natural selection in the human lineage. *Science* **312**(5780), 1614–1620 (2006).

Sangster, T. A., Lindquist, S., and Queitsch, C. Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. *Bioessays* **26**(4), 348–362 (2004).

Schacherer, J., Shapiro, J. A., Ruderfer, D. M., and Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of saccharomyces cerevisiae. *Nature* **458**(7236), 342–345 (2009).

Schluter, D. Adaptive radiation along genetic lines of least resistance. *Evolution* **50**(5), 1766–1774 (1996).

Schmalhausen, I. I. *Factors of Evolution: the Theory of Stabilizing Selection*. The Blakiston Company, Philadelphia, PA, (1949).

Schwenk, K. and Wagner, G. P. Function and the evolution of phenotypic stability: connecting pattern to process. *American Zoologist* **41**, 552–563 (2001).

Shalgi, R., Lapidot, M., Shamir, R., and Pilpel, Y. A catalog of stability-associated sequence elements in 3′ UTRs of yeast mRNAs. *Genome Biol* **6**(10), R86 (2005).

Shmulevich, I., Kauffman, S. A., and Aldana, M. Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proc Natl Acad Sci U S A* **102**(38), 13439–13444 (2005).

Sia, R. A., Bardes, E. S., and Lew, D. J. Control of Swe1p degradation by the morphogenesis checkpoint. *EMBO J* **17**(22), 6678–6688 (1998).

Sia, R. A., Herald, H. A., and Lew, D. J. Cdc28 tyrosine phosphorylation and the morphogenesis checkpoint in budding yeast. *Mol Biol Cell* **7**(11), 1657–1666 (1996).

Simola, D. F. and Kim, J. Evolutionary dynamics of the *Saccharomyces* genome and a resource for population genomics. In revision., (2008).

Slatkin, M. Quantitative genetics of heterochrony. *Evolution* **41**(4), 799–811 (1987).

Smith, J. M., Burian, R., Kauffman, S., Alberch, P., Campbell, J., Goodwin, B., Lande, R., Raup, D., and Wolpert, L. Developmental constraints and evolution: a perspective from the mountain lake conference on development and evolution. *The Quarterly Review of Biology* **60**(3), 265–287 September (1985).

Smith, J. M. and Haigh, J. The hitch-hiking effect of a favorable gene. *Genetics Research* **23**, 23–25 (1974).

Sniegowski, P. D., Dombrowski, P. G., and Fingerman, E. *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res* **1**(4), 299–306 (2002).

Somel, M., Franz, H., Yan, Z., Lorenc, A., Guo, S., Giger, T., Kelso, J., Nickel, B., Dannemann, M., Bahn, S., Webster, M. J., Weickert, C. S., Lachmann, M., Paabo, S., and Khaitovich, P. Transcriptional neoteny in the human brain. *Proc Natl Acad Sci U S A* **106**(14), 5743–5748 (2009).

Soranzo, N., Zampieri, M., Farina, L., and Altafini, C. mRNA stability and the unfolding of gene expression in the long-period yeast metabolic cycle. *BMC Syst Biol* **3**(18) February (2009).

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **9**(12), 3273–3297 (1998).

Stajich, J. E. and Hahn, M. W. Disentangling the effects of demography and selection in human history. *Mol Biol Evol* **22**(1), 63–73 (2005).

Stearns, S. C. *The Evolution of Life Histories*. Oxford University Press, New York, NY, (2004).

Stearns, S. C. and Magwene, P. The naturalist in a world of genomics. *Am Nat* **161**(2), 171–180 (2003).

Stern, D. L. Perspective: Evolutionary developmental biology and the problem of variation. *Evolution* **54**(4), 1079–1091 August (2000).

Strobeck, C., Smith, J. M., and Charlesworth, B. The effects of hitchhiking on a gene for recombination. *Genetics* **82**(3), 547–558 (1976).

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* **39**(12), 1443–1452 (2007).

Sweeney, J. Y., Kuehne, H. A., and Sniegowski, P. D. Sympatric natural *Saccharomyces cerevisiae* and *S. paradoxus* populations have different thermal growth profiles. *FEMS Yeast Res* **4**(4-5), 521–525 (2004).

Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 November (1989).

Tanay, A., Regev, A., and Shamir, R. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A* **102**(20), 7203–7208 (2005).

Tawfik, O. W., Papasian, C. J., Dixon, A. Y., and Potter, L. M. *Saccharomyces cerevisiae* pneumonia in a patient with acquired immune deficiency syndrome. *J Clin Microbiol* **27**(7), 1689–1691 July (1989).

Systat Software, Inc. *SigmaPlot Version 11*. Systat Software, Inc., San Jose, CA, (2007).

Wolfram Research, Inc. *Mathematica Edition: Version 5.0*. Wolfram Research, Inc., Champaign, Il, (2003).

Tibshirani, R., Walther, G., and Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B.* **63**(2), 411–423 (2001).

Tirosh, I., Reikhav, S., Levy, A. A., and Barkai, N. A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324**(5927), 659–662 (2009).

Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol* **4**, 159 (2008).

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**(1), 31–40 (2007).

Townsend, J. P., Cavalieri, D., and Hartl, D. L. Population genetic variation in genome-wide gene expression. *Mol Biol Evol* **20**(6), 955–963 (2003).

Tsai, I. J., Bensasson, D., Burt, A., and Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc Natl Acad Sci U S A* **105**(12), 4957–4962 (2008).

Vanoni, M., Vai, M., and Frascotti, G. Effects of temperature on the yeast cell cycle analyzed by flow cytometry. *Cytometry* **5**(5), 530–533 (1984).

Vilela, C. and McCarthy, J. E. G. Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. *Mol Microbiol* **49**(4), 859–867 August (2003).

von Linné, C. *Systema Naturae per Regna Tria Naturae, Secundum Classes, Ordines, Genera, Species, cum Charateribus, Differentiis, Synonymis, Locis*. Holmiae: Impensis Direct, (1758).

Waddington, C. H. Evolution of developmental systems. *Nature* **147**, 108–110 (1941).

Waddington, C. H. Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563–565 (1942).

Waddington, C. H. *The Evolution of an Evolutionist*. Cornell University Press, Ithaca, NY, (1975).

Wagner, G. P., Booth, G., and Bagheri-Chaichian, H. A population genetic theory of canalization. *Evolution* **51**(2), 329–347 (1996).

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., and Brown, P. O. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci U S A* **99**(9), 5860–5865 (2002).

Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**(2), 256–276 (1975).

Wickens, M., Bernstein, D. S., Kimble, J., and Parker, R. A PUF family portrait: 3′ UTR regulation as a way of life. *Trends Genet* **18**(3), 150–157 (2002).

Willmore, K. E., Young, N. M., and Richtsmeier, J. T. Phenotypic variability: its components, measurement and underlying developmental processes. *Evolutionary Biology* **34**, 99–120 (2007).

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**(5429), 901–906 (1999).

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat Genet* **40**(3), 346–350 (2008).

Yuan, Y., Guo, L., Shen, L., and Liu, J. S. Predicting gene expression from sequence: a reexamination. *PLoS Comput Biol* **3**(11), e243 (2007).

Zeyl, C. and DeVisser, J. A. Estimates of the rate and distribution of fitness effects of spontaneous mutation in *Saccharomyces cerevisiae*. *Genetics* **157**(1), 53–61 (2001).

Zhang, Z. and Dietrich, F. S. Identification and characterization of upstream open reading frames (uORF) in the 5′ untranslated regions (UTR) of genes in *Saccharomyces cerevisiae*. *Curr Genet* **48**(2), 77–87 (2005).

Zhao, H., Butler, E., Rodgers, J., Spizzo, T., Duesterhoeft, S., and Eide, D. Regulation of zinc homeostasis in yeast by binding of the ZAP1 transcriptional activator to zinc-responsive promoter elements. *J Biol Chem* **273**(44), 28713–28720 (1998).